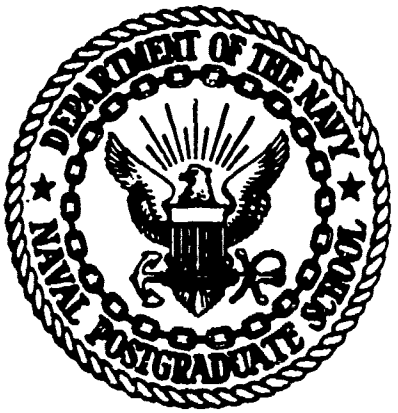
NAVAL POSTGRADUATE SCHOOL
Monterey, California



THESIS

A FREQUENCY DOMAIN ANALYSIS OF
THE LINEAR DISCRETE KALMAN FILTER

by

Walter Jeremiah Costello

March 1980

Thesis Advisor: R.W. Hamming

Approved for public release; distribution unlimited.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A085 058 / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Frequency Domain Analysis of the Linear Discrete Kalman Filter.	5. TYPE OF REPORT & PERIOD COVERED Master's Thesis; March 1980
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Walter Jeremiah Costello	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Postgraduate School Monterey, California 93940	12. REPORT DATE Mar[illegible]80
	13. NUMBER OF PAGES ~~100~~ 12 100
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Kalman Filter
Digital Filters
Smoothing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The linear discrete Kalman filter was analyzed using a frequency domain approach. Process and measurement noise covariances are shown to be critical design parameters which, together with the assumed prior state and covariance estimates, completely determine the gain schedule of the linear Kalman filter. Several relevant design techniques are illustrated and discussed. The concepts of smoothing and sharpening are

DD FORM 1 JAN 73 1473 (Page 1) EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

251450

demonstrated. Extensions to adaptive, non-linear, and non-parametric filtering are briefly discussed, as are applications to inventory management, estimation of time-varying mean functions, and multiple regression.

Approved for public release; distribution unlimited

A Frequency Domain analysis of the Linear Discrete Kalman Filter

by

Walter Jeremiah Costello
Lieutenant Colonel, United States Marine Corps
B.E.E., Rensselaer Polytechnic Institute, 1963

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
March 1980

Author Walter J. Costello

Approved by: Richard W. Hamming
Thesis Advisor

D. P. Gaver
Second Reader

James K Hartman FOR M. G. SOVEREIGN
Chairman, Department of Operations Research

K. T. Schrady
Dean of Information and Policy Sciences

ABSTRACT

The linear, discrete Kalman filter was analyzed using a frequency-domain approach. Process and measurement noise covariances are shown to be critical design parameters which, together with the assumed prior state and covariance estimates, completely determine the gain schedule of the linear Kalman filter. Several relevant design techniques are illustrated and discussed. The concepts of smoothing and sharpening are demonstrated. Extensions to adaptive, non-linear, and non-parametric filtering are briefly discussed, as are applications to inventory management, estimation of time-varying mean functions, and multiple regression.

TABLE OF CONTENTS

I. INTRODUCTION ---------- 8

A. BACKGROUND ---------- 8

B. PURPOSE ---------- 10

C. METHOD ---------- 10

D. LEVEL OF PRESENTATION ---------- 10

E. SUMMARY OF RESULTS ---------- 11

II. THEORY ---------- 13

A. STOCHASTIC PROCESSES AND STATIONARITY ---------- 13

B. THE PHILOSOPHICAL CONCEPT OF STATIONARITY ------- 14

C. DUALITY OF THE TIME AND FREQUENCY DOMAINS ------- 15

1. Fourier Series ---------- 15

2. Basic Concept ---------- 17

3. Discrete Data and the Sampling Theorem ------ 17

4. The Discrete Fourier Transform ---------- 18

D. THE DOOB-MEYER-FISK DECOMPOSITION ---------- 20

E. DIGITAL FILTERS ---------- 22

1. Some Classifications of Digital filters ----- 23

2. Applications of Digital Filters ---------- 24

3. Analysis of Digital Filters ---------- 24

F. DATA ANALYSIS AND EXPERIMENT DESIGN ---------- 25

III. THE LINEAR DISCRETE KALMAN FILTER ---------- 27

A. DESCRIPTION ---------- 27

B. THE SCALAR KALMAN FILTER ---------- 31

1. Transient and Steady-State Gain ---------- 31

2. Frequency Response ---------- 32

C. IMPROVING THE FILTER ---------- 36

D. PERFORMANCE COMPARISON ---------- 37

E. THE TRANSIENT CASE ---------- 48

F. HIGHER ORDER FILTERS ---------- 50

IV. ESTIMATION, SMOOTHING AND PREDICTION ---------- 62

A. ESTIMATION ---------- 62

B. SMOOTHING ---------- 62

C. A COMPARISON OF TWO SMOOTHERS ---------- 65

D. PREDICTION ---------- 69

V. SOME REFINEMENTS, EXTENSIONS AND ALTERNATIVES ------- 75

A. ADAPTIVE FILTERING ---------- 75

B. NON-LINEAR FILTERING ---------- 78

1. Non-Linear Measurements ---------- 78

2. Non-Linear Dynamics ---------- 79

C. NON-PARAMETRIC FILTERING ---------- 82

VI. SOME APPLICATIONS ---------- 86

A. INVENTORY MANAGEMENT ---------- 86

B. ESTIMATING A MEAN FUNCTION ---------- 87

C. MULTIPLE REGRESSION ---------- 89

D. SOME DESIGN CONSIDERATIONS ---------- 90

1. Spectral Analysis of the Data ---------- 90

2. Frequency Analysis of Proposed Models ------- 91

3. Adjusting the Model ---------- 91

4. Testing the Model ---------- 92

APPENDIX A. DERIVATIONS ---------- 93

BIBLIOGRAPHY ---------- 97

INITIAL DISTRIBUTION LIST ---------- 99

LIST OF FIGURES

1. Gibbs Phenomenon ---------- 17
2. Linear Discrete Kalman Filter ---------- 29
3. Amplitude Response of Scalar Kalman Filter --------- 34
4. Phase Shift of Scalar Kalman Filter ---------- 35
5. Impulse Response of Scalar Filters ---------- 38
6. Amplitude Comparison of Single and Double Filter --- 39
7. Phase Shift of Single and Double Filter ---------- 40
8. Experimental Data ---------- 42
9. Spectral Analysis of Data ---------- 43
10. Single and Double Filter Performance ---------- 44
11. Single and Double Filter Performance ---------- 46
12. Single and Double Filter Performance ---------- 47
13. Gain Schedule Comparison ---------- 49
14. Velocity Filter Impulse Response ---------- 53
15. Frequency Response of Position Estimate ---------- 54
16. Frequency Response of Velocity Estimate ---------- 55
17. Velocity Filter Performance ---------- 57
18. Velocity Filter Performance ---------- 59
19. Frequency Response of Acceleration Filter ---------- 61
20. Kalman Smoother Performance ---------- 64
21. Amplitude Response of Kalman and Gaussian Smoothers- 67
22. Filter Weights, Kalman and Gaussian Smoothers ------ 68
23. Gaussian Smoother Performance ---------- 70
24. Comparison of Kalman and Gaussian Smoothers -------- 71
25. Comparison of Kalman and Gaussian Smoothers -------- 72
26. Adaptive Filters ---------- 77

I. INTRODUCTION AND SUMMARY

A. BACKGROUND

The Kalman filter is a recursive Bayesian least-squares estimator of an n-dimensional system state vector based on an m-dimensional measurement vector. The filter may operate in a J-dimensional coordinate system where J≥m, J≥n. The basic assumption is that each dimension of the coordinate system varies according to a kth order Gauss-Markov process. The Kalman Filter was developed in the early 1960's by Kalman and Bucy [refs. 1 and 2].

The Kalman filter may be used to obtain an optimal estimate of the present state, a prediction of future states, and/or smoothed estimates of past states. The current state estimate is generally used to determine an optimal control input. Future state estimates are used to determine optimum present policy. Smoothed past state estimates are used for data analysis and model building. Thus the potential areas of application span the field of time series analysis.

Applications of the Kalman filter are numerous and the theory is being continually developed and extended. An overview of the development of linear filtering theory and an extensive bibliography may be found in Kailath [ref.3]. A reasonably clear presentation of theory and applications is contained in Gelb [ref.4].

Perhaps the widest and most successful application of Kalman filtering has been to vehicle tracking and control. Clark [ref.5] has written a particularly lucid description of the design of a filter for an anti-aircraft gun fire control system which is noteworthy for its clarity of presentation of the underlying theory. It is evident [refs. 4 and 5] that the design process is heuristic, and requires extensive testing and analysis of candidate filter configurations, even when the process is well-understood and is based on a mature technology.

The Kalman filter has also been applied, with varying degrees of success, in economic models, inventory models, and even weather models. Considerable difficulty is encountered in model building, because the filter design requires good estimates of the variance and covariance of noise sources, as well as an accurate state transition model. A prior estimate of system state and covariance is also required, which is somewhat less critical because errors in the prior estimate decrease with time. These parameters are often difficult to determine in highly random processes of questionable stationarity.

The Kalman filter is derived and designed almost entirely within the time domain, although Clark [ref.5] does refer to the concept of filter bandwidth. The Kalman filter is essentially a low-pass filter with a very wide transition band, and higher-order filters have some amplification at the mid or low-mid frequency range. In general, the stop band does not completely attenuate high frequencies. This allows

the filter to attenuate high-frequency noise somewhat while still retaining some response to sudden changes of state.

B. PURPOSE

The purpose of this thesis is to acquaint the reader with the Kalman filter, to show how the choice of various filter parameters affect its performance, and to provide design insight through analysis in the frequency domain. The approach is tutorial, and the reader is referred to some of the interesting examples which may be found in the literature.

C. METHOD

The frequency response of several simple filter designs were investigated using the Fast Fourier Transform program in the APL Library 2. The computer results were justified analytically for the simplest design, a scalar single-state filter. Derivations are presented in appendix A.

D. LEVEL OF PRESENTATION

Full understanding of the theory requires a knowledge of stochastic processes that evolve over time, as well as an understanding of digital signal theory in the frequency domain. The Fourier transform is a basic tool. A full exposition of the underlying theory is clearly beyond the

scope of this presentation. The reader is directed to Larson and Shubert [ref.6] for the theory of stochastic processes and to Hamming [ref.7] for the theory of digital filtering. As previously mentioned, Gelb [ref.4] and Clark [ref.5] are good references for the Kalman filter. Bloomfield [ref.8] and Brillinger [ref.9] are also applicable references. Brown [ref.10] and Box and Jenkins [ref.11] contain related material.

There are few readers who are entirely conversant with both the frequency domain and time domain approach to time series analysis. Nevertheless, a duality exists between the two, and a summary of the theory is presented.

Illustrative examples will often be based on tracking models, because this is presently the widest area of application of Kalman filters, and because most readers will find the concepts of position, velocity, and acceleration easy to understand. The concepts are easily extendable to other areas. For example, the economist may wish to replace "velocity" with "trend".

E. SUMMARY OF RESULTS

The steady-state gain, bandwidth, and sensitivity of the linear discrete Kalman filter are shown to be completely determined by the choice of the process and measurement noise covariances. Filter performance on stationary or nearly-stationary data can be predicted by comparing the frequency response of the proposed filter with a spectral

analysis of the data. The wide transition band of the amplitude response of the scalar Kalman filter can be sharpened by multiple passes of the data through a higher-gain filter. This can be accomplished simply and recursively. The superiority of symmetric smoothing filters over non-symmetric filters was demonstrated. When used as a smoother (by using both forward and backward passes) the Kalman filter was as effective as a non-recursive Gaussian filter. Higher-order filters were shown to have higher bandwidth and amplification as the order of the filter was increased. A frequency domain approach to filter design may provide additional insight and enable the designer to achieve better filter performance, particularly when the system state transition model and noise covariance models are not well understood.

II. THEORY

A. STOCHASTIC PROCESSES AND STATIONARITY

A continuous stochastic process X(t) is a Gaussian process if the probability densities of all orders are multivariate Gaussian densities. It is a kth order Gauss-Markov process if the state at time t depends only on k earlier states. If we should expand the state space to k states, which include all derivatives up to the (k-1)th, the future system state vector will depend only on the present state. For example, if the acceleration of a vehicle is a first-order Gauss-Markov process, then the position of the vehicle is a third-order Gauss-Markov process. However, if our state space includes acceleration and velocity as well as position, the future state of the system is independent of all but the present state. If the random acceleration has zero mean, and variance one over one time increment, the acceleration is a standard Wiener process W(t). The derivative of the Wiener process, written dW(t), has zero mean, unit variance, and is called white Gaussian noise, which may be thought of as a "zero-th order" Gauss-Markov process [ref.4].

The standard Wiener process is not stationary, because the variance grows linearly with time. That is, the estimate of a future state based on the present state has variance

that is a linear function of time. However, the standard Wiener process has stationary, independent increments. That is, the variance at time (t+1) given the state at time (t) is constant and independent of t.

A stochastic process X(t) is wide-sense stationary if and only if it has a constant mean function, and a correlation function such that [ref.6]

$$R_x(t_1 + s, t_2 + s) = R(t_1, t_2) = R(t_2 - t_1)$$

that is, the correlation function of the process is independent of an arbitrary time shift s. A Gaussian process is strictly stationary if and only if it is wide-sense stationary [ref.6].

The Gauss-Markov assumption makes possible the development of theory and applications, because, in general, any linear operation performed on a Gaussian process results in another Gaussian process, and the Markov property allows consideration of only the present state, disregarding all previous states.

B. THE PHILOSOPHICAL CONCEPT OF STATIONARITY

A frequency-domain analysis of a stochastic process is only meaningful if the process is stationary. If the process were changing over time, the spectrum would change over time. Since the spectrum can only be analyzed by means of data taken over time, such analysis of a non-stationary process

would be meaningless. However, if the process is "quasi-stationary", that is, it exhibits stationary statistics for a while, then undergoes a change, then settles down to stationarity again, the frequency approach is still useful, although inaccurate over the transition period. As an example, consider an airplane subject to random accelerations due to air turbulence. An appropriate model might be a third-order Gauss-Markov process as long as the airplane maintains a straight path or turns at a constant acceleration. However, the pilot's inputs to initiate or terminate a maneuver would result in brief periods of non-stationarity, and the model would perform inadequately during and immediately after the transition period.

It may be argued that every practical process can be considered stationary over infinite time. If the process is random, it represents an ensemble of possible paths, of which any realization in terms of real-world data is only one possible path, and may or may not be closely representative of the ensemble. When dealing with reality, we are often forced to assume stationarity in order to make analysis possible, and often we obtain good results even though we can never know whether or not the assumption of stationarity is really valid.

C. DUALITY OF THE TIME AND FREQUENCY DOMAINS

1. Fourier Series

A very wide class of mathematical functions may be

represented by the Fourier series [ref.12] as follows:

$$s(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nt + b_n \sin nt)$$

where $0 \leq t \leq 2\pi$.

Existence and convergence of this series representation require only that $s(t)$ be everywhere single-valued, and possess a finite number of maxima, minima, and finite discontinuities. The function $s(t)$ need not be differentiable. Any function meeting the above criteria can be thought of as a constant mean function a_0, plus an infinite series of sines and cosines of integral frequencies and various amplitudes. Of course, the independent variable t must be shifted and scaled to the interval $[\, 0, 2\pi \,]$. Note that the lowest frequency present, aside from the zero-frequency mean, is one cycle for the span of $s(t)$. Among the functions meeting the criteria are a square pulse, an impulse, and any manifestation of a random walk. In practice, the Fourier analysis of a function $s(t)$ requires the truncation of the infinite Fourier series. This results in a smooth least-squares approximation to the function $s(t)$. There are ripples in the approximation if the function $s(t)$ is not differentiable or if the truncation is too severe. This is known as the Gibbs phenomenon, and is illustrated in figure 1, which was taken from Hamming [ref.7]. By taking a sufficient number of terms in the Fourier expansion, we can improve the closeness of the approximation.

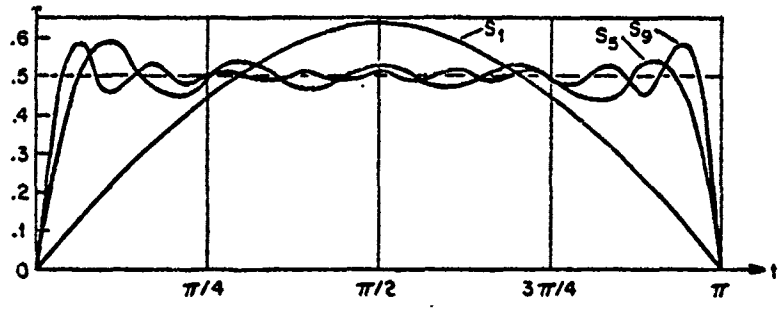


Figure 1. The Gibbs Phenomenon

2. Basic Concept

The basic concept of the duality between the time and frequency domains is so simple that it often gets lost in a forest of Fourier transforms. The time period is the reciprocal of frequency. The basic relationship is

$$v/2\pi = f = 1/T$$

where v is the frequency in radians/unit time, f is the frequency in cycles/unit time, and T is the time period for one cycle. Stated simply, frequency is the inverse of the time period.

3. Discrete Data and the Sampling Theorem

The digital computer allows the efficient analysis of continuous phenomena by means of discrete approximations. We saw earlier that the lowest frequency contained in a Fourier expansion of a function g(t) was the reciprocal of the time span covered by the function. Similarly, the famous Sampling Theorem [refs. 6 and 7] states that if a function g(t) in continuous time is sampled at constant, discrete time

intervals Δt (that is, at a rate of $1/\Delta t$), then the highest observable frequency is 0.5 cycles per measurement interval Δt. This means that at least two observations are required in each cycle in order to observe that particular frequency. The frequency $0.5/\Delta t$, usually written simply 0.5, is referred to as the Nyquist frequency. The result is the aliasing phenomenon, which is familiar to most moviegoers. During the chase, the stagecoach wheels appear to stop or rotate slowly backwards when the rate of rotation of the wheel spokes (spokes/sec) exceeds 1/2 the camera rate (frames/sec). When higher frequencies exist in the function $g(t)$ sampled at a rate Δt, they are folded back and appear in the frequency spectrum of the sampled data as frequencies less than the Nyquist frequency. The sampling theorem shows that a spectral analysis of discrete data is only meaningful over the Nyquist interval $[-0.5/\Delta t, +0.5/\Delta t]$.

4. The Discrete Fourier Transform

Any function $g(t)$ for which a convergent Fourier series exists may be represented in the frequency domain in terms of real and imaginary parts, or in terms of amplitude and phase angle, as a function of frequency. It should be noted that the function $g(t)$ may also be complex-valued, but we will deal with only real-valued functions. In the continuous domain, the formulas

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(v) \exp(ivt)\, dv$$

and

$$G(v) = \int_{-\infty}^{\infty} g(t)\ \exp(-ivt)\ dt$$

represent a Fourier transform pair. The frequency response G(v) completely determines the time function g(t) and conversely.

If the function g(t) is sampled at intervals t = 0,1,2,...,n the time-to-frequency transformation becomes

$$G(v) = \sum_{t=0}^{n} g(t)\ \exp(-ivt) \qquad t = 0,1,...,n$$

which is defined only on the Nyquist interval $[-\pi, \pi]$, here defined in radians. The formula may be written in a more familiar form by using the Euler relation

$$\exp(-ivt) = \cos vt - i \sin vt$$

as

$$G(v) = \sum_{t=0}^{n} g(t)\ (\cos vt - i \sin vt)$$

which is continuous in v on the interval $[-\pi, \pi]$. The Fourier transform is a bit difficult to handle analytically for all but the simplest functions, but the discrete Fourier transform is generally easy to compute by use of a Fast Fourier Transform (FFT) program available in most computer libraries. The output will generally be a very close discrete approximation to G(v), if the span of g(t) is large enough. The inverse transformation can also be made.

Since G(v) is complex valued whenever the function g(t) is not symmetric, it is often useful to represent it in terms of amplitude and phase. The amplitude is

$$|G(v)| = \sqrt{G(v)G(-v)} = \sqrt{[Re(v)]^2 + [Im(v)]^2}$$

where Re(v) and Im(v) are the real and imaginary parts of G(v). The phase angle is

$$\theta(v) = \arctan Im(v) / Re(v)$$

D. THE DOOB-MEYER-FISK DECOMPOSITION

In most practical applications, a finite-variance sample-continuous stochastic process X(t) can be written

$$X(T) = X(0) + \int_0^T A(t)dt + \int_0^T B(t)dW(t)$$

where X(0) is the initial value of that process, A(t)dt is predictable, smooth behavior determined by a set of deterministic differential equations describing the system, and B(t)dW(t) is noise, where dW(t) is white Gaussian noise, and B(t) is a smooth transformation that is sometimes thought of as "coloring" the noise. Such a representation is called the Doob-Meyer-Fisk decomposition [ref.6], which may be thought of as separating the process into a signal and noise. Several important points must be made with regard to this

equation. It is not intended here that the expression be evaluated analytically. The integral $B(t)dW(t)$ is an Itō integral, which is not even a stochastic version of a Stieltjes integral [ref.6]. Also, although the processes $A(t)$ and $B(t)$ are smooth functions that may be considered deterministic representations of system behavior, they are not necessarily known to the observer, even when an adequate technological representation exists.

Consider again our piloted aircraft being tracked by a radar. The process $A(t)$ represents the dynamics of the airframe, as affected by the control inputs of the pilot, which are unknown to the radar observer. The process $B(t)$ consists of several parts. One is the measurement process, which may or may not be known to the radar observer. For example, a nutating radar antenna might impose some periodic error in the measurement, which would be manifested in the process $B(t)$. Overlaid on this might be a white Gaussian noise measurement error. Air turbulence could also be represented as white Gaussian noise, which, however, could only be manifested through deterministic airplane dynamics. There are those who would argue that the pilot should also be modelled as a random variable. In any event, the process $B(t)$ might be further decomposed into several processes, here at least airframe response to air turbulence and periodic radar antenna dynamics.

The vital observation is that if the frequency content of the processes $A(t)$ and $B(t)$ are known to be different, they can be partially separated by a spectral analysis of the

data. In our example, aircraft have natural dynamic response frequencies in all control axes. These can be estimated closely, even for enemy airplanes, and are generally similar among similar types of airplanes, although they vary with airspeed. It is physically impossible for the airplane to respond faster than its highest natural dynamic frequencies. Any frequency content higher than this must be noise. If the radar system dynamics are of a higher frequency than this, they can also be separated. The pilot will take advantage of the full response rate of the airplane only very rarely. Therefore, low frequency components are most likely due to pilot maneuvers. Of course, since white Gaussian noise has a flat frequency spectrum as a result of aliasing [ref.7], it is impossible to separate all of the noise from the signal. However, it is often possible to remove quite a bit of it.

E. DIGITAL FILTERS

A digital filter is a linear transformation applied iteratively to a set of data points. The purpose here is to separate noise from the signal. The simplest digital filter is the simple average, which estimates the mean value from the data, and smooths out all fluctuations. The most general form of the digital filter was stated by Hamming [ref.7] as

$$x(t) = \sum_{k=-\infty}^{\infty} a(k)\, z(t-k) + \sum_{k=1}^{\infty} b(k)\, x(t-k)$$

where the estimate $x(t)$ at some point t is a linear

combination of the data points z(t-k), and perhaps of the previous estimates x(t-k). The coefficients a(k) and b(k) are weighting coefficients and may, of course, be zero. As a result of the sampling theorem, the filtering process is meaningless unless the measurements z(t-k) are made at equally spaced intervals along the t axis, where t is usually, but not necessarily, time.

1. Some Classifications of Digital Filters

Digital filters may be classified as symmetric or non-symmetric, and as recursive or non-recursive. A symmetric non-recursive filter is one in which all b(k) equal zero and all a(k) = a(-k), such as the filter

$$x(t) = 0.2\ z(t-1) + 0.6\ z(t) + 0.2\ z(t+1).$$

An example of a recursive filter is

$$x(t) = a\ z(t) + b\ x(t-1) \quad 0<a<1,\ b = 1-a$$

which is not symmetric. This particular filter may be expressed as

$$x(t) = a\ z(t) + b[\ a\ z(t-1) + b[\ a\ z(t-2) + \ldots.]]$$

which reduces to

$$x(t) = a\ z(t) + ab\ z(t-1) + ab^2 z(t-2) + \ldots + ab^n z(t-n) + \ldots$$

The recursive filter extends to the infinite past, although the coefficients ab^n will approach zero, if $|b|<1$. In this case, a recursive filter can be closely approximated by a non-recursive filter. A primary advantage of the recursive filter is that old data need not be stored. New estimates may be computed simply and rapidly as time evolves. This is an important advantage for real-time applications.

2. Applications of Digital Filters

Digital filters are used to separate a signal from noise, to separate various frequency components of a signal, and/or to perform such mathematical functions as integration and differentiation. A review of Simpson's rule and the Trapezoidal rule should convince the reader that these numerical integration techniques are, in fact, recursive digital filters. Sometimes a filter has two purposes. For example, it might be desirable, n estimating velocity from successive observations of position, to simultaneously differentiate and remove high frequency noise. When a filter is used to stop part of the frequency spectrum, it is referred to as a "low-pass", "high-pass", "band-pass", or "band-stop" filter, depending on its function.

3. Analysis of Digital Filters

In the time domain, a digital filter is described completely by its impulse response function, which is nothing

more than the response of the filter to data consisting of a string of zeros, a single one, followed by zeros. The output of the filter is then simply the weighting coefficients a(k). If the filter is recursive, we might not be able to deduce the recursive form from the coefficients a(k), but that will not concern us here. The Fourier transform of the impulse-response function

$$H(v) = \sum_{k=-N}^{N} a(k) \exp(-ivk)$$

will completely specify the frequency response of the filter. If the filter is symmetric, there will be no imaginary part, and hence no phase shift. If the filter is recursive, it cannot practically be symmetric, and the summation will generally run from zero to infinity. That is, the impulse response will extend infinitely far into the future, which means that the filter remembers all of the past.

The duality of the time and frequency domains allows us to specify a desired frequency response and to design an appropriate filter by calculating filter weights, or to analyze an existing filter by calculating the frequency response from the filter weighting coefficients.

F. DATA ANALYSIS AND EXPERIMENT DESIGN

No digital filter should be applied to data analysis without a clear idea of the effect of the filter upon the data. Slutsky and Yule first noted that some smoothing

formulas induced periodic functions in the smoothed estimate that were more the effect of smoothing than of the original data [ref.7]. A spectral analysis of representative raw data can be helpful in deciding on an appropriate filtering technique. However, such data as economic time series or weather data are typically very noisy, are based on a relatively short run of data, and cannot be described by an adequate technological model. The analyst must be aware of these problems. Sometimes there are no good solutions, but a spectral representation may produce frequencies that can be explained on rational grounds.

Another potential pitfall is a result of the sampling theorem. Consider the timely example of an air pollution model. It would be reasonable to suspect that air pollution would follow at least a daily cycle, or perhaps an eight hour cycle if morning and evening rush hours were considered. Daily samples of air pollution could not hope to uncover cycles of a shorter period than every two days. Samples every four hours would be marginally adequate. Hourly samples would be necessary for a good analysis. Additionally, recall the requirement for equally-spaced sampling intervals. For various reasons, the analyst may have no control over data collection. However, he must always understand what has been done, or could have been done, to the data, as well as what he is doing to it, in order to avoid erroneous conclusions.

III. THE LINEAR DISCRETE KALMAN FILTER

A. DESCRIPTION

The linear discrete Kalman filter is a recursive Bayesian least-squares estimator of the state vector of a linear system based on a vector of noisy measurements made at discrete time intervals. The process to be estimated is assumed to be an n-state Gauss-Markov process of order k, subject to process noise W with zero mean and covariance matrix Q. The process is observed by an m-dimensional measurement, subject to measurement noise V (not to be confused with frequency (v)) with zero mean and covariance matrix R. The filter requires a prior Bayesian estimate of system state and covariance. The recursive estimate of system state at time t is obtained by the formula

$$X(t|t) = X(t|t-1) + K(t)[Z(t) - HX(t|t-1)]$$

where

X(t|t)	state estimate based on current measurement
Z(t)	current measurement
X(t|t-1)	state estimate prior to current measurement
K(t)	Kalman gain matrix (to be discussed later)
H	observation matrix, which is constant

The derivation of the Kalman filter equations may be

found in Gelb [ref.4]. A summary of the filter equations is presented in figure 2, which should be consulted in order to follow the subsequent discussion.

In general, the state model represents a dynamic system, that is, one which changes with time. The extrapolation of the state estimate to the time of the next observation is obtained by the formula:

$$X(t+1|t) = \Phi X(t|t)$$

where Φ is the state transition matrix. The observation process occurs according to the conceptual relation

$$Z(t) = HX(t) + V$$

where X(t) is the true system state observed through the observation matrix H, and V represents measurement noise, which is assumed to be a Gaussian random variable with zero mean and covariance matrix R. Note that the process represented by this formula is assumed to occur in the real world. The computation does not occur in the filter. Rather, the measurement Z(t) is an input to the filter.

In the linear Kalman filter, the gain K(t) does not depend in any way on the data. It depends only on the model, and is therefore extremely sensitive to assumptions. Gain is calculated according to the formula

$$K(t) = P(t)H^T[HP(t)H^T + R]^{-1}$$

LINEAR DISCRETE KALMAN FILTER

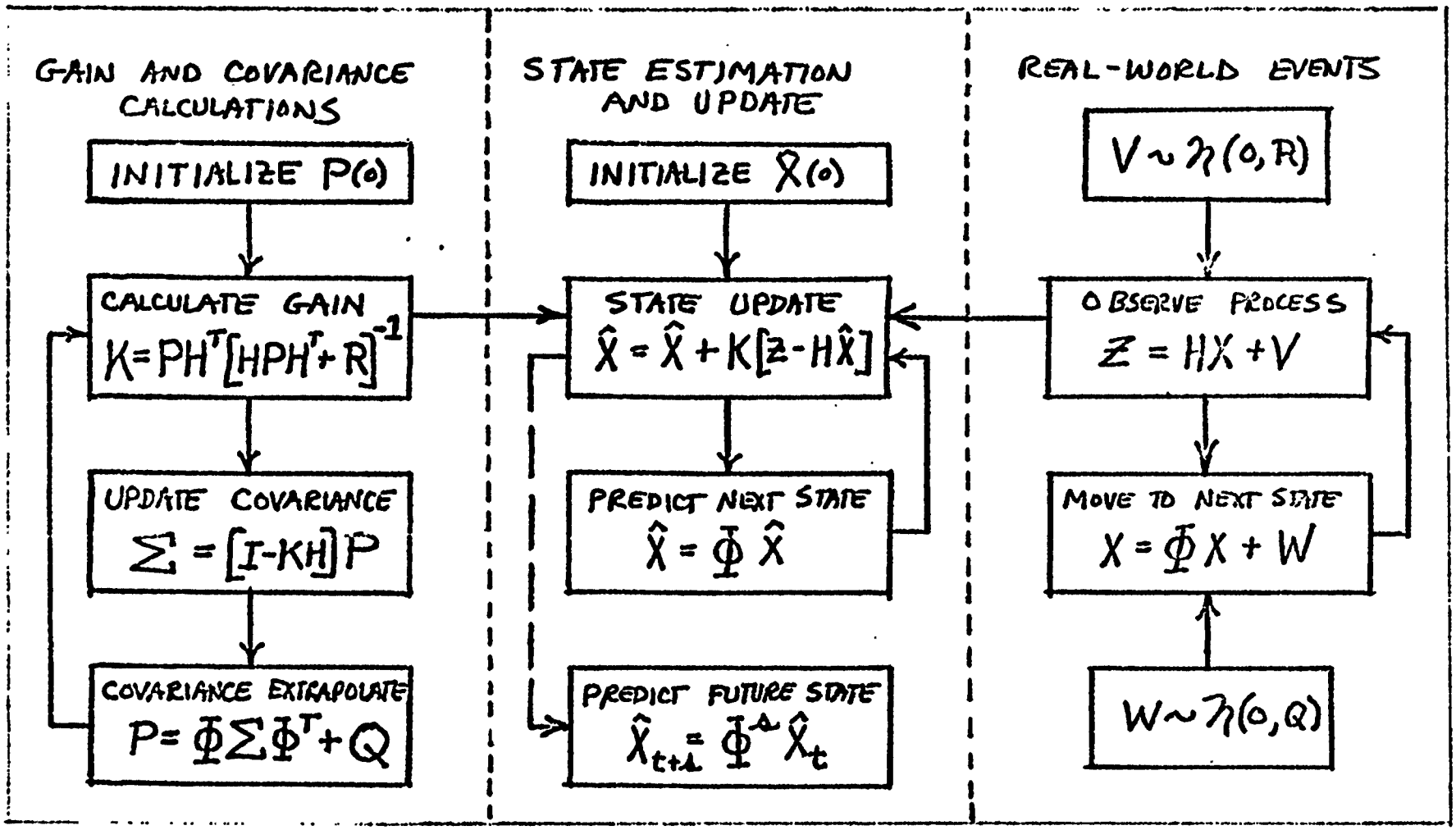


FIGURE 2

where P(t) is the covariance in the system state estimate prior to the current measurement and R is the covariance of the measurement error. The covariance is updated according to the formula

$$\Sigma(t) = [I-K(t)H]\ P(t)$$

where $\Sigma(t)$ is the state covariance given the current measurement, and I is the identity matrix. The covariance is extrapolated to the time immediately prior to the next observation by the formula

$$P(t+1) = \Phi\ \Sigma(t)\ \Phi^T + Q$$

where Φ is the state transition matrix and Q is the covariance of the process noise. Combining the above two equations shows that the covariance of the state estimate at the time of the current measurement depends on the previous covariance according to the formula

$$\Sigma(t) = [I-K(t)H]\ [\Phi\ \Sigma(t-1)\ \Phi^T + Q]$$

Filter performance is very dependent on adequate modelling, particularly on the state transition model Φ and the choice of noise covariances R and Q. To a lesser extent, performance also depends on the initial estimates of system state X(1|0) and covariance P(1|0). However, the latter parameters are less important because their effects decrease

with time. If the matrices R, Q, Φ, and H are constant in time, the gain K(t) and covariance matrices $\Sigma(t)$ and P(t) eventually reach a steady state, and are completely determined by R, Q, Φ, and H.

For a given linear filter, it will be shown that filter gain, covariance, and frequency response will be completely determined by the choice of R and Q.

B. THE SCALAR KALMAN FILTER

The multi-state Kalman filter is a powerful computational device. However, it is difficult and often impossible to manipulate in closed form because of the frequent occurence of singular matrices. An analysis of a single-state (scalar) filter can be used to illustrate the mechanics of the Kalman filter, and to aid in developing an intuitive understanding. In the discussion that follows, it is assumed that all matrices are scalars, and, in particular, Q and H equal one. Matrix notation is preserved for clarity. Derivations may be found in appendix A.

1. Transient and Steady-State Gain

It can be shown (appendix A) that the scalar Kalman gain can be expressed recursively as

$$K(t) = \frac{K(t-1) + Q/R}{K(t-1) + Q/R + 1}$$

When the filter reaches steady-state, the gain is constant

and

$$K = \frac{-Q}{2R} + \sqrt{\frac{Q^2}{4R^2} + \frac{Q}{R}}$$

The inverse relationship is

$$\frac{Q}{R} = \frac{K^2}{1-K}$$

Thus, the variance ratio Q/R, which is the ratio of process noise variance to measurement noise variance, completely determines the steady-state gain. The steady-state filter is completely described by the formula

$$X(t) = K\ Z(t) + (1-K)\ X(t-1)$$

2. Frequency Response

Letting $K = a$ and $(1-K) = b$, the impulse-response function G(t) may be written

$$G(t) = ab^t,\ t=0,1,2,.....$$

The Fourier Transform is

$$H(v) = \int_{-\infty}^{\infty} G(t)\ \exp(-ivt)dt$$

$$H(v) = a \sum_{t=0}^{\infty} [b\ \exp(-iv)]^t$$

$$H(v) = a\ /\ [1 - b\ \exp(-iv)]$$

Since the filter is not symmetric, the frequency response H(v) has both real and imaginary parts. The amplitude may be written

$$A = |H(v)| = \sqrt{H(v)H(-v)} = a / \sqrt{1 + b^2 - 2b \cos v}$$

which reduces to

$$A = \sqrt{\frac{Q/R}{Q/R + 2(1 - \cos v)}}$$

The phase angle may be written

$$\theta(v) = \arctan\left(\frac{-b \sin v}{1 - b \cos v}\right)$$

The angle for maximum phase shift is

$$v(\max \theta) = \arccos b = \arccos (1-K)$$

$$v(\max \theta) = \arccos (1 + Q/2R - \sqrt{Q^2/4R^2 + Q/R})$$

Therefore, the variance ratio Q/R also completely specifies the steady-state frequency response of the filter. Amplitude and phase relationship for several values of gain are plotted in figures 3 and 4.

It is evident that high Q/R (high gain) reduces the phase lag of the filter but allows more of the high-frequency components to pass. Conversely, low Q/R (low gain) attenuates more of the high-frequency components, at the expense of an increased phase lag. Note that even at very

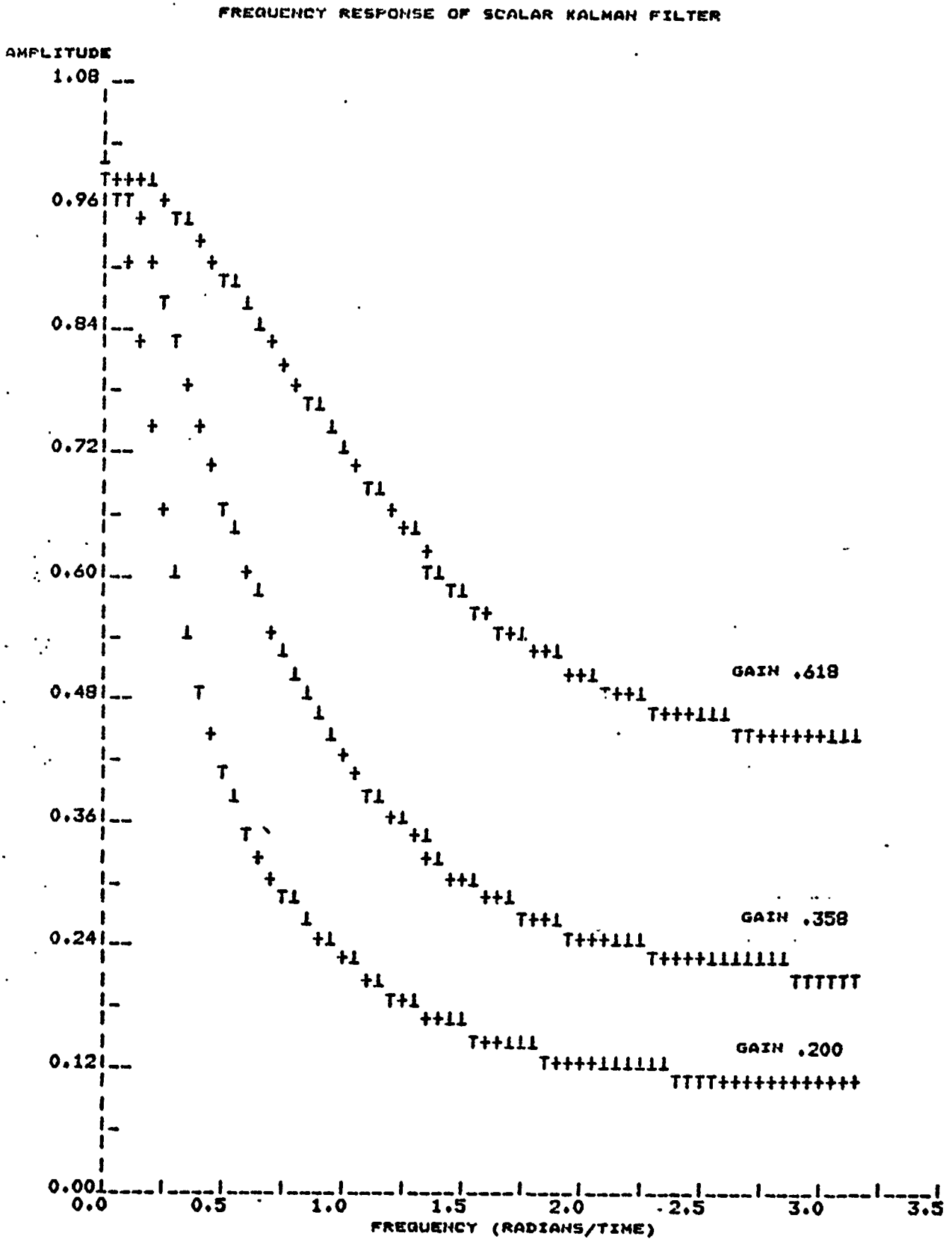

FREQUENCY RESPONSE OF SCALAR KALMAN FILTER
AMPLITUDE
1.08
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
GAIN .618
GAIN .358
GAIN .200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 3

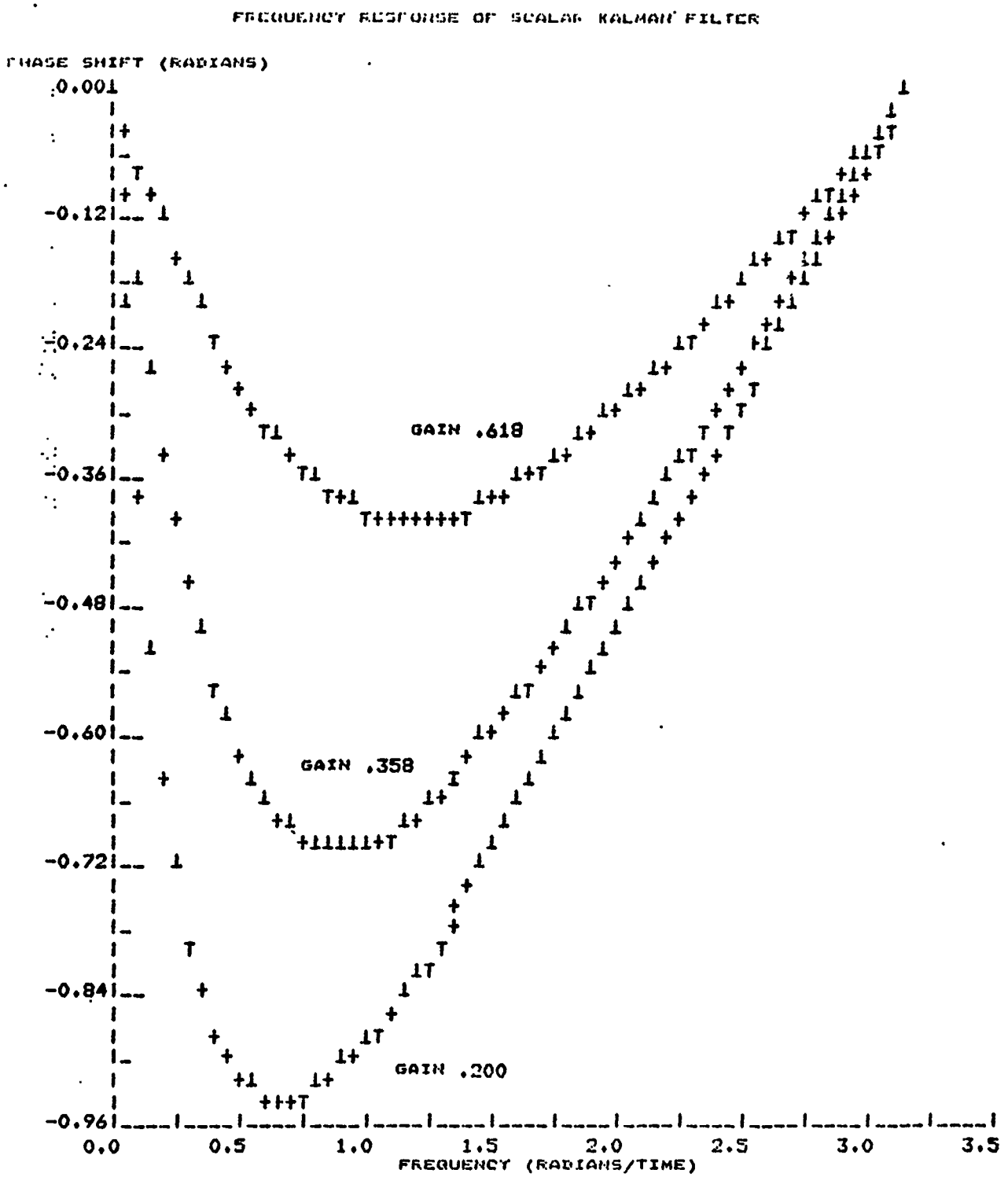

FREQUENCY RESPONSE OF SCALAR KALMAN FILTER
PHASE SHIFT (RADIANS)
0.001
-0.12
-0.24
-0.36
-0.48
-0.60
-0.72
-0.84
-0.96
GAIN .618
GAIN .358
GAIN .200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 4

low gain (low Q/R) not all of the high-frequency component is attenuated, and the phase lag is quite severe. The slope of the amplitude change is quite shallow, implying that attenuation increases gradually as frequency increases. This is a consequence of the assumptions and performance will not be adequate if the data does not represent a Gauss-Markov process, but in fact represents some phenomena changing with time.

C. IMPROVING THE FILTER

The transition band of the filter can be sharpened, and more of the high- frequency components eliminated, by running the data through two filters in series. The basic scalar filter was

$$x(t) = a\ z(t) + b\ x(t-1)$$

where a = K and b = (1-K). Running the data through the filter again, we obtain a new estimate y(t) , where

$$y(t) = a\ x(t) + b\ y(t-1)$$

It should be evident that we can accomplish this all in one step as

$$y(t) = a^2 z(t) + 2b\ y(t-1) - b^2 y(t-2)$$

We need only to save one additional previous estimate y(t-2) as well as y(t-1). The impulse response function is

$$g(t) = (t+1)a^t b^t, \quad t = 0,1,...$$

We have performed a convolution in the time domain, which corresponds to a multiplication in the frequency domain. This may not be exactly what we want. Let us suppose that we want the weighting coefficient for the present data point z(t) to be 0.27 in both cases. This requires a = 0.27 for the basic filter and a = $\sqrt{0.27}$ = 0.52 for the double filter. The impulse response function for both filters is presented in figure 5. Note that the double filter forgets the past more readily. The amplitude and phase shift for both filters is presented in figures 6 and 7. The gain for the scalar filter was 0.27, corresponding to a variance ratio (Q/R) of 0.1. Also note that the double filter has somewhat better high-frequency attenuation, somewhat less attenuation at low frequencies, and a slightly sharper (steeper) transition. We would therefore expect it to be a bit better at separating a low-frequency signal from noise. At low frequencies, the double filter has less phase shift. However, phase shift is more severe at frequencies above 0.5 radians.

D. PERFORMANCE COMPARISON

The basic scalar filter and the improved (double) filter

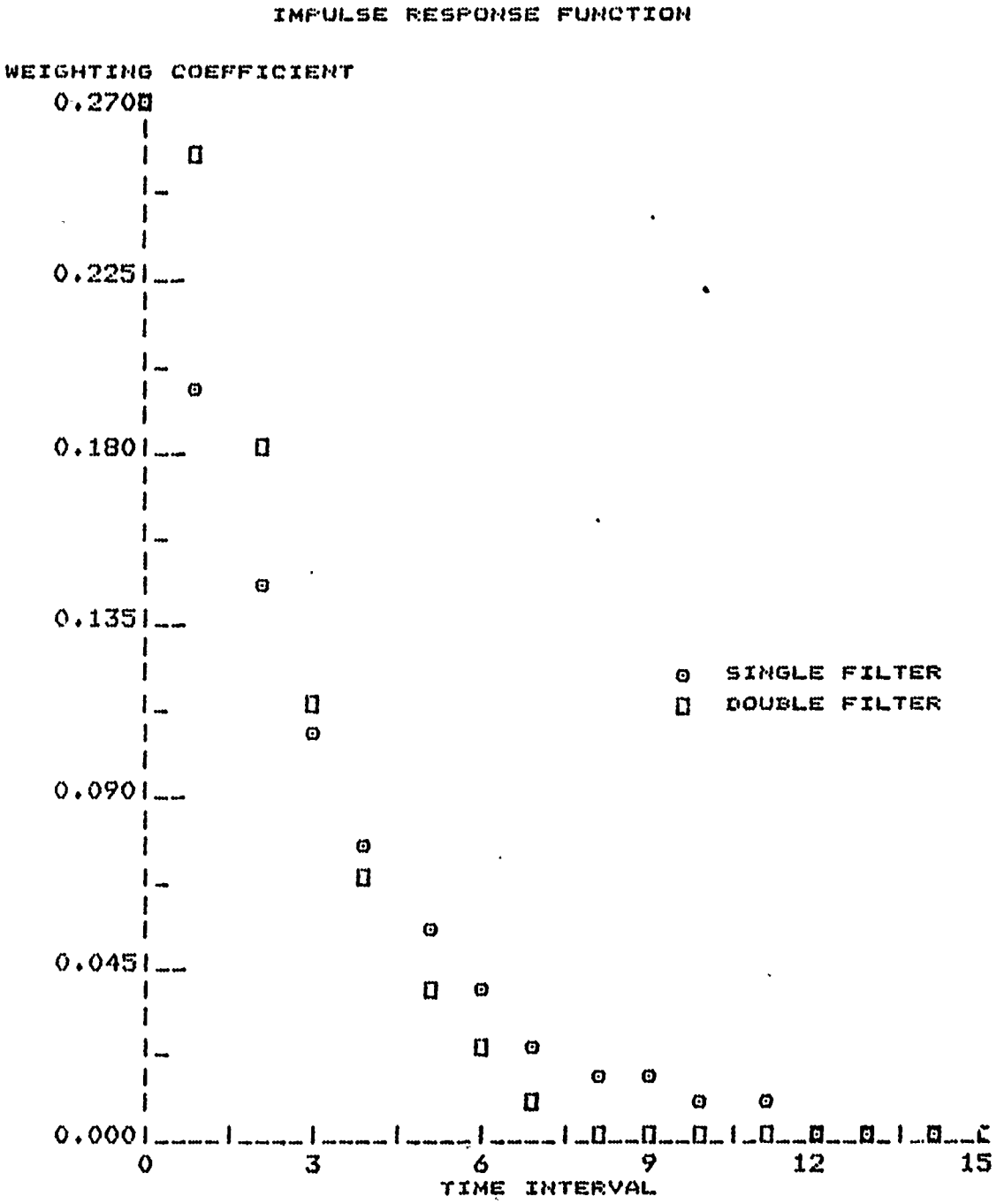

IMPULSE RESPONSE FUNCTION
WEIGHTING COEFFICIENT
0.270
0.225
0.180
0.135
0.090
0.045
0.000
0
3
6
9
12
15
TIME INTERVAL
SINGLE FILTER
DOUBLE FILTER

FIGURE 5

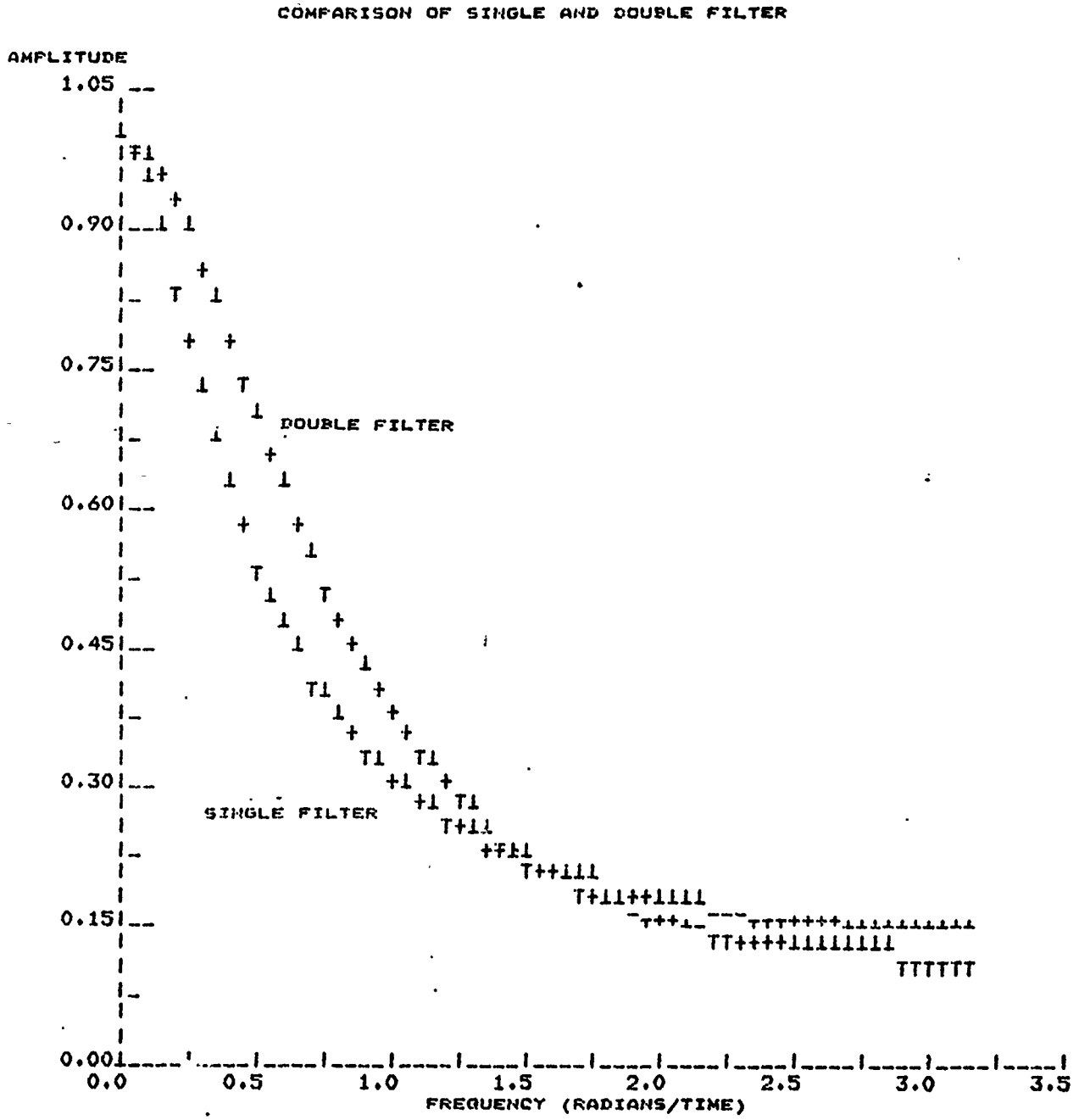

COMPARISON OF SINGLE AND DOUBLE FILTER
AMPLITUDE
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00
DOUBLE FILTER
SINGLE FILTER
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 6

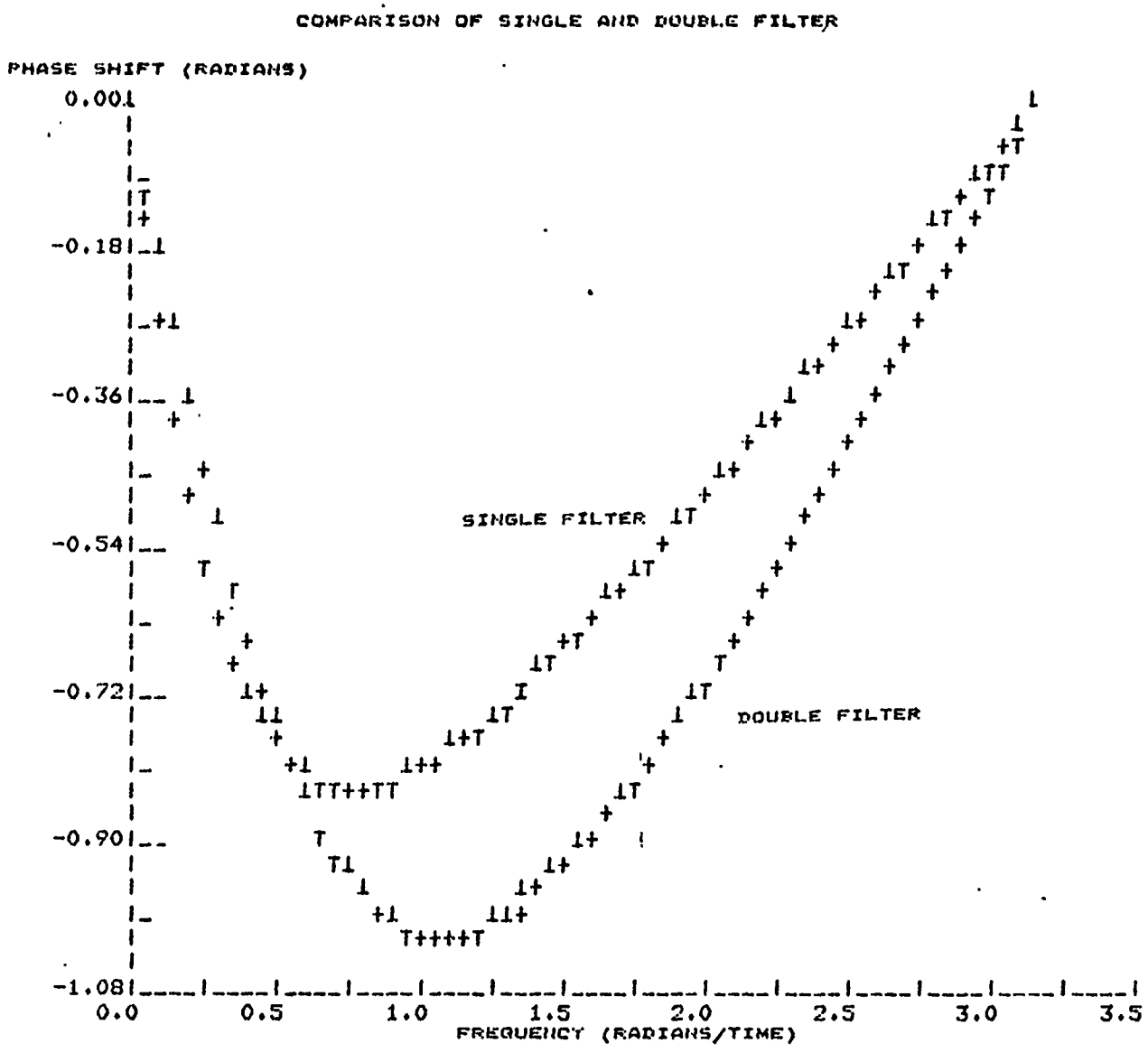

COMPARISON OF SINGLE AND DOUBLE FILTER
PHASE SHIFT (RADIANS)
0.00
-0.18
-0.36
-0.54
-0.72
-0.90
-1.08
SINGLE FILTER
DOUBLE FILTER
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 7

were compared using the data illustrated in figure 8. A series of independent normal random numbers with zero mean and variance ten were generated, to simulate measurement noise. They were added to a nominal function which was a combination of steps and a ramp. The ramp function rises 1 unit each measurement interval, which corresponds to one standard deviation in process noise Q, so at least during the ramp function, the data corresponds to the filter design variance ratio Q/R = 0.1. A spectral analysis of the data is presented in figure 9. The signal and noise are presented separately and in combination. Note that the noise-only scale is expanded. The noise spectrum is irregular, but overall quite flat. The signal consists mostly of very low frequencies, but also has some high frequencies. This would be expected, since step and ramp functions require very high frequencies in their Fourier expansion. The high-frequency signal is submersed in noise. The data, of course, does not fulfil the assumptions from which the Kalman filter is derived. However, the real world seldom does either. We are looking for robustness.

The filters were first tested on the signal alone. The results are presented in figure 10. It can be seen that neither filter can respond instantaneously to the discontinuities in the function, since high frequencies are attenuated. Both filters lag after discontinuities and during the ramp rise. This is a consequence of the non-symmetric nature of the filters and illustrates the phase lag. Note that the double filter performs a bit better,

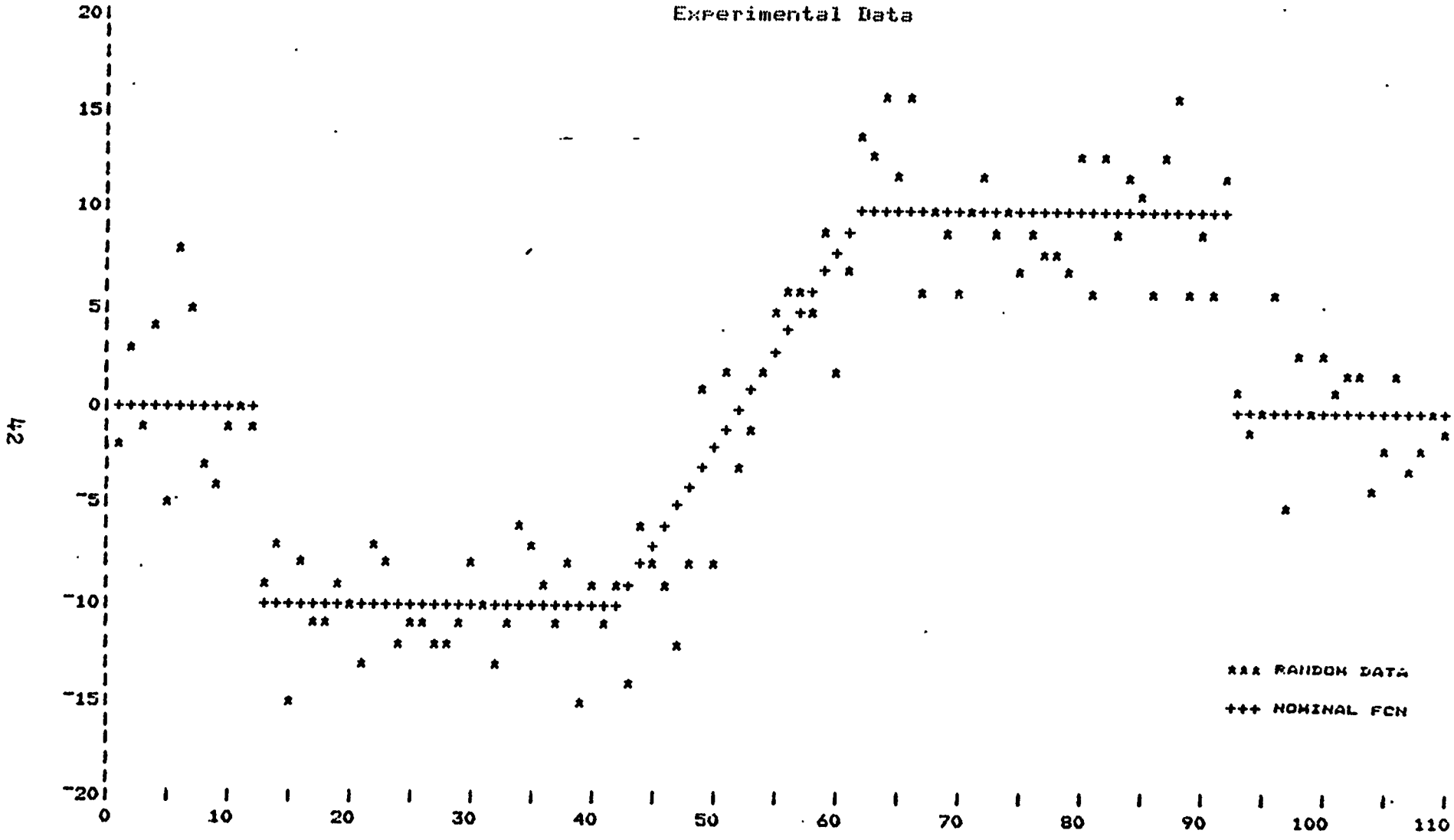

Experimental Data
20
15
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50
60
70
80
90
100
110
*** RANDOM DATA
+++ NOMINAL FCN

FIGURE 8

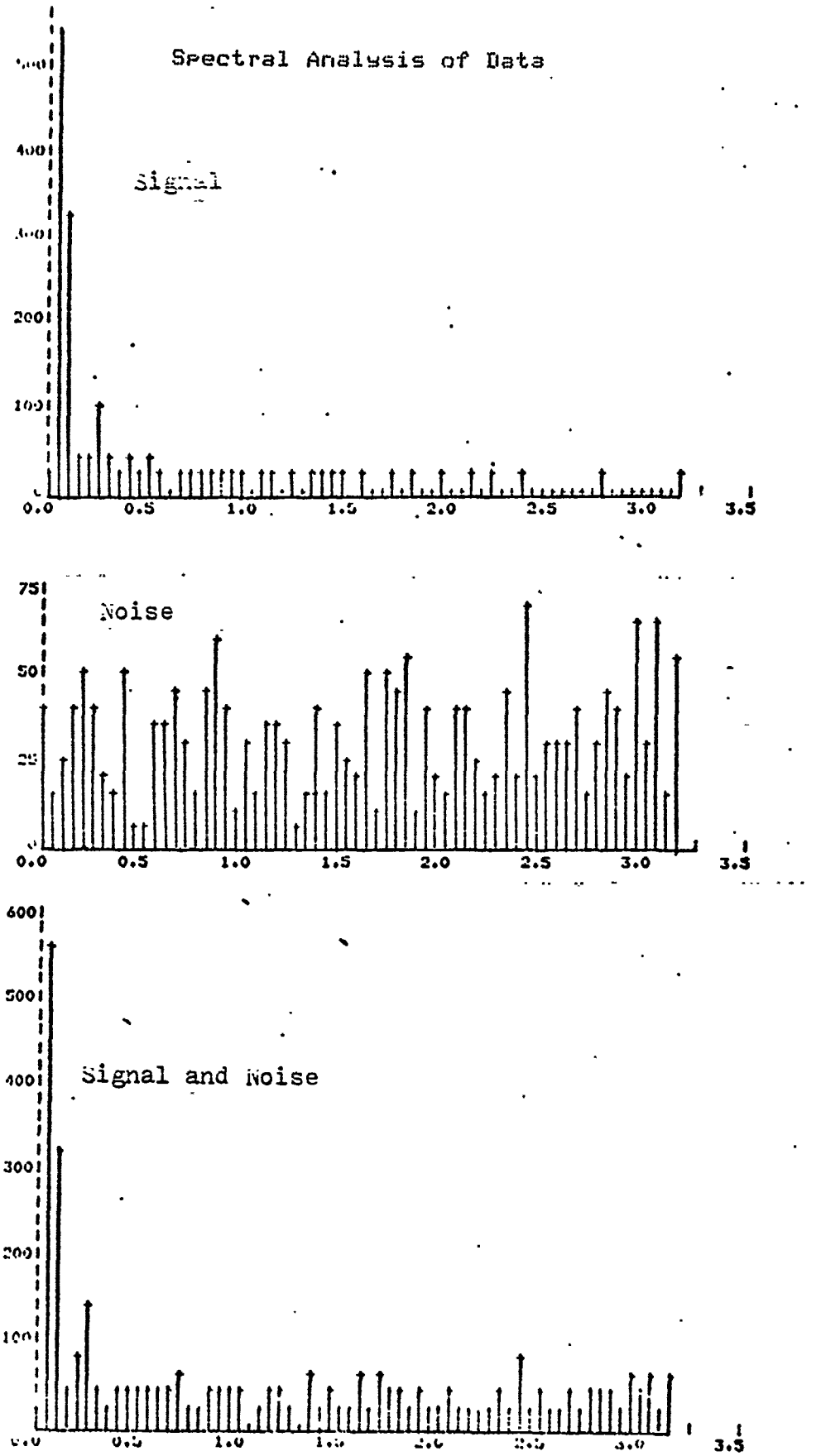

Spectral Analysis of Data
Signal
400
200
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Noise
75
50
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
600
500
400
Signal and Noise
300
200
100
0.0
0.5
1.0
3.0
3.5

FIGURE 9

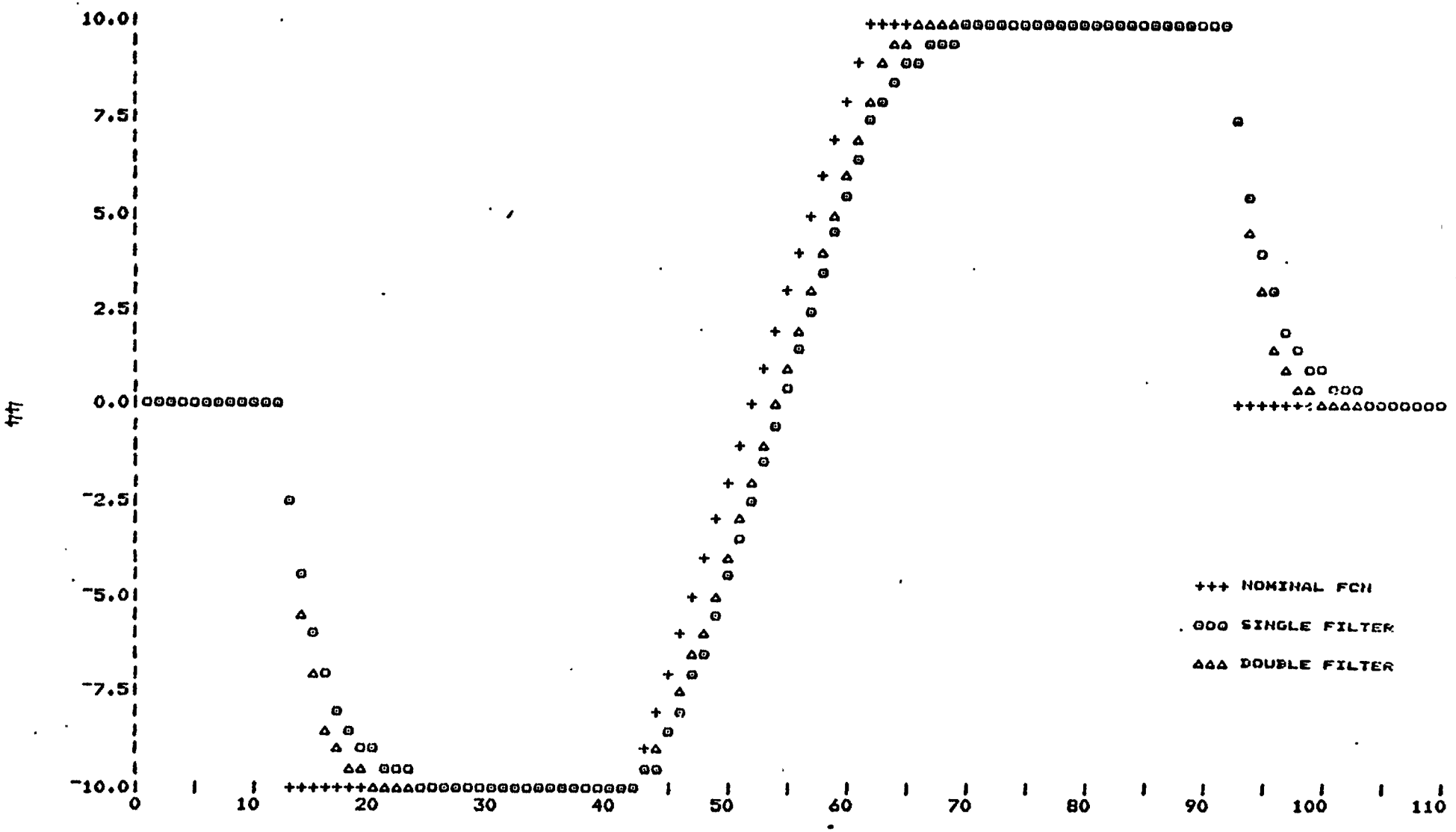

10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10.0
0
10
20
30
40
50
60
70
80
90
100
110
+++ NOMINAL FCN
ooo SINGLE FILTER
ΔΔΔ DOUBLE FILTER

FIGURE 10

remaining closer to the data throughout (it is a peculiarity of the plotting routine that only one symbol will be plotted when ever data points coincide. Therefore, data points that do not appear should be regarded as occurring simultaneously with the ones that do appear).

The filters were then tested on the noisy data. The results are presented in figures 11 and 12. The latter plot has the data suppressed so that the scale can be increased and more resolution obtained.

The same trends can be observed as were previously. The double filter lags less during the trend and transitions. The double filter appears to follow the noise a bit more closely, but overall it follows the signal better than the basic filter. The average variance between the signal and the filter was 4.20 for the double filter and 5.10 for the basic filter, which was an 18% improvement for this simple modification. The improvement is due to the fact that the double filter weights more recent data more heavily, and remembers less of the past than the scalar Kalman filter, even though the weight on the present observation is the same.

This simple experiment is only intended to acquaint the reader with possible improvements to the Kalman filter. Like any tool, the Kalman filter should not be applied indiscriminantly. The interested reader is referred to Hamming [ref.7] for the basic principles of digital filter design.

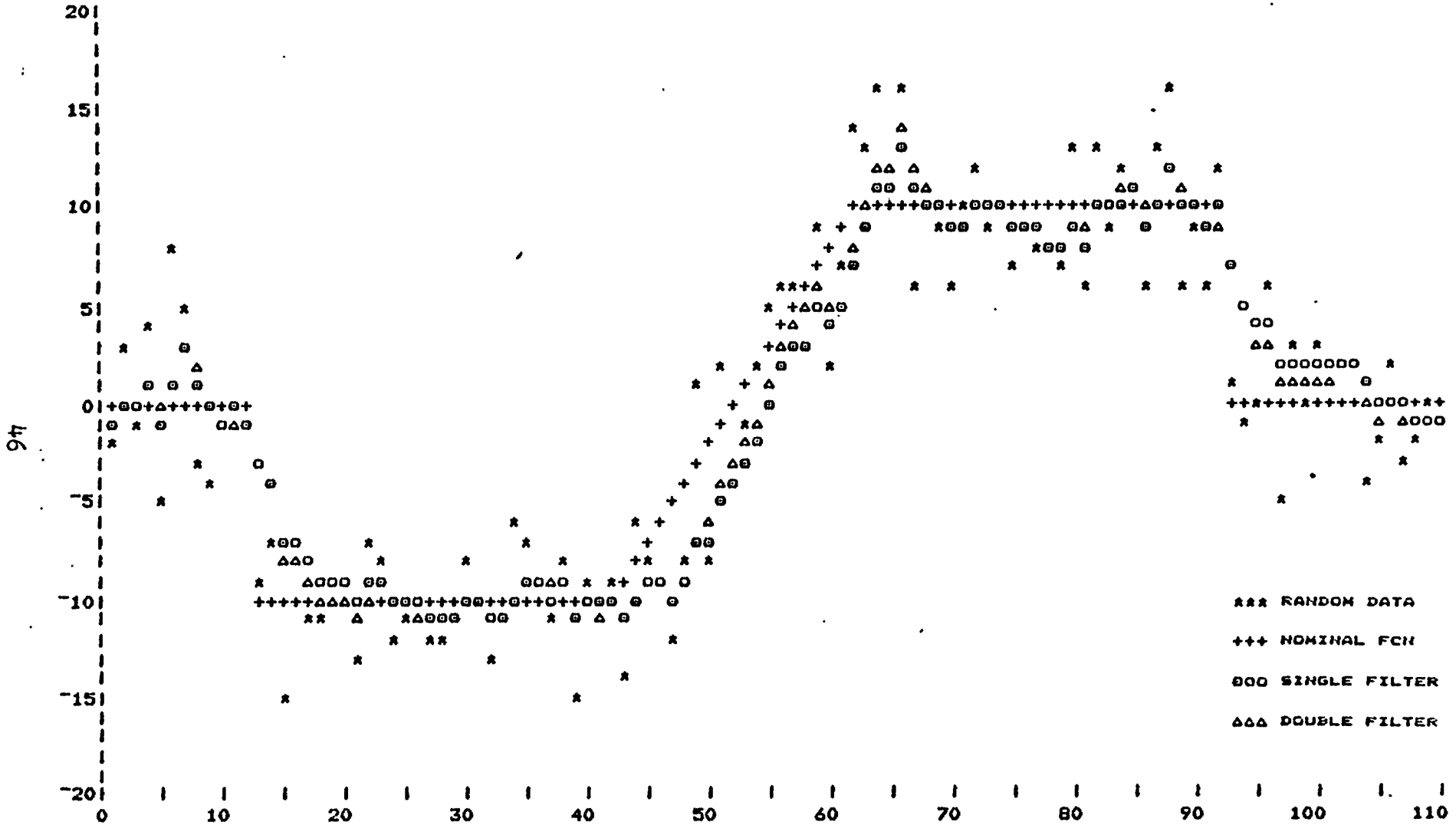

20
15
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50
60
70
80
90
100
110
η
*** RANDOM DATA
+++ NOMINAL FCN
ooo SINGLE FILTER
ΔΔΔ DOUBLE FILTER

FIGURE 11

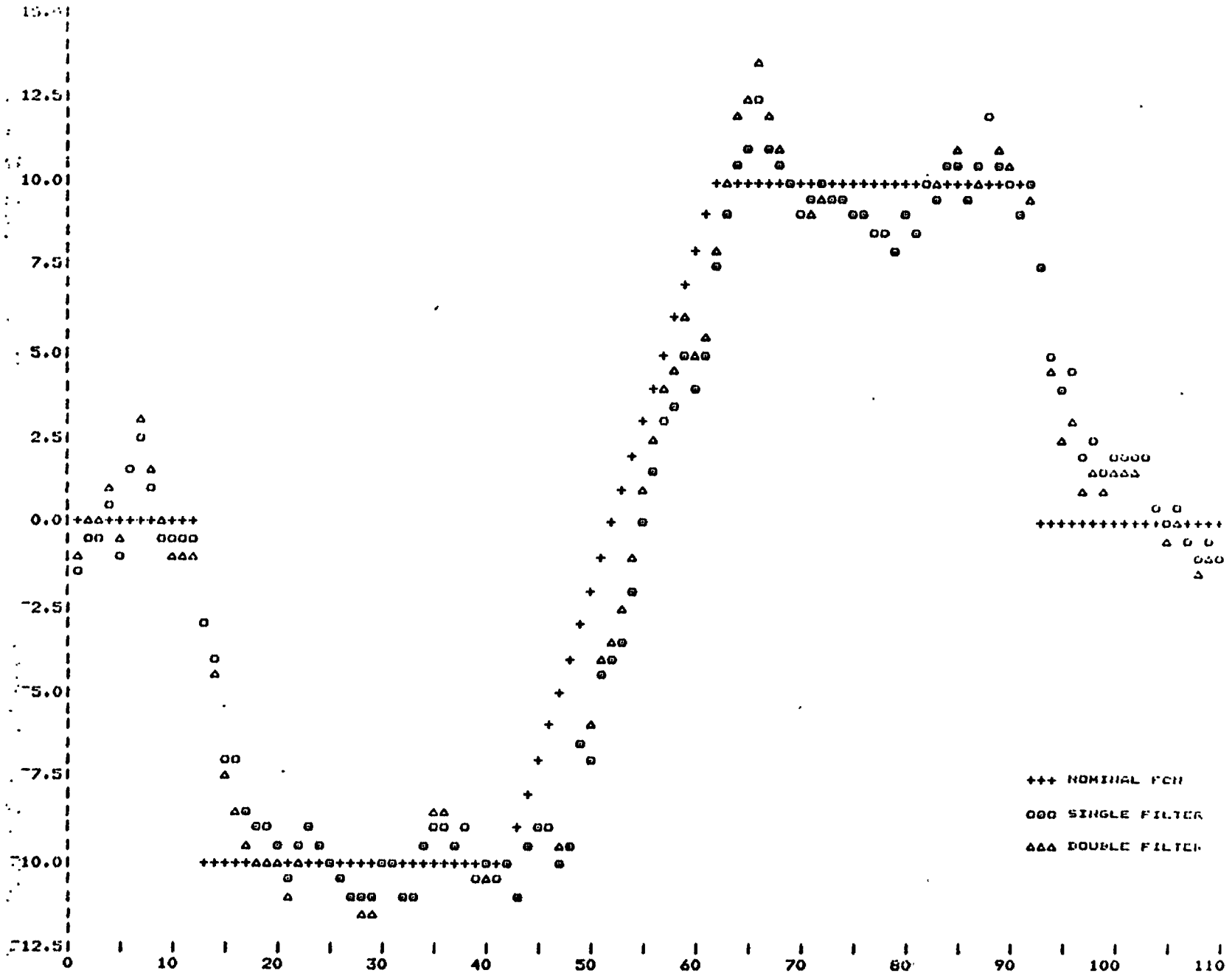

15.0
12.5
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10.0
-12.5
0
10
20
30
40
50
60
70
80
90
100
110
+++ NOMINAL FCN
ooo SINGLE FILTER
ΔΔΔ DOUBLE FILTER

FIGURE 12

E. THE TRANSIENT CASE

The sophisticated reader has no doubt noticed that the steady-state scalar Kalman filter is equivalent to the Box-Jenkins IMA(0,1,1) model [ref.11] and to Brown's exponential smoothing model [ref.10]. Zehna criticized the exponential smoothing model [ref.13], noting the bias would occur if the steady-state model was applied with an inappropriate prior estimate. Bessler and Zehna [ref.14] developed a gain schedule which they call finite exponential smoothing. Their formula for gain is

$$a(t) = a/(1-b^t)$$

where a is the steady state gain, b = (1-a), and a(t) is the gain schedule as a function of time. It is similar to the Kalman gain schedule if the initial Kalman gain K(0) is chosen as one. In both models, no prior estimate is required. The weight on the first observation is one. A comparison of the two models is illustrated in figure 13, for a steady-state gain of 0.2 and an initial gain of 1. The Kalman gain was calculated according to the recursive formula in section III.A.

The Kalman filter gain converges faster, although the difference is not great. The scalar Kalman filter possesses two other advantages over the finite exponential smoothing technique. First, if a good prior estimate does exist, the

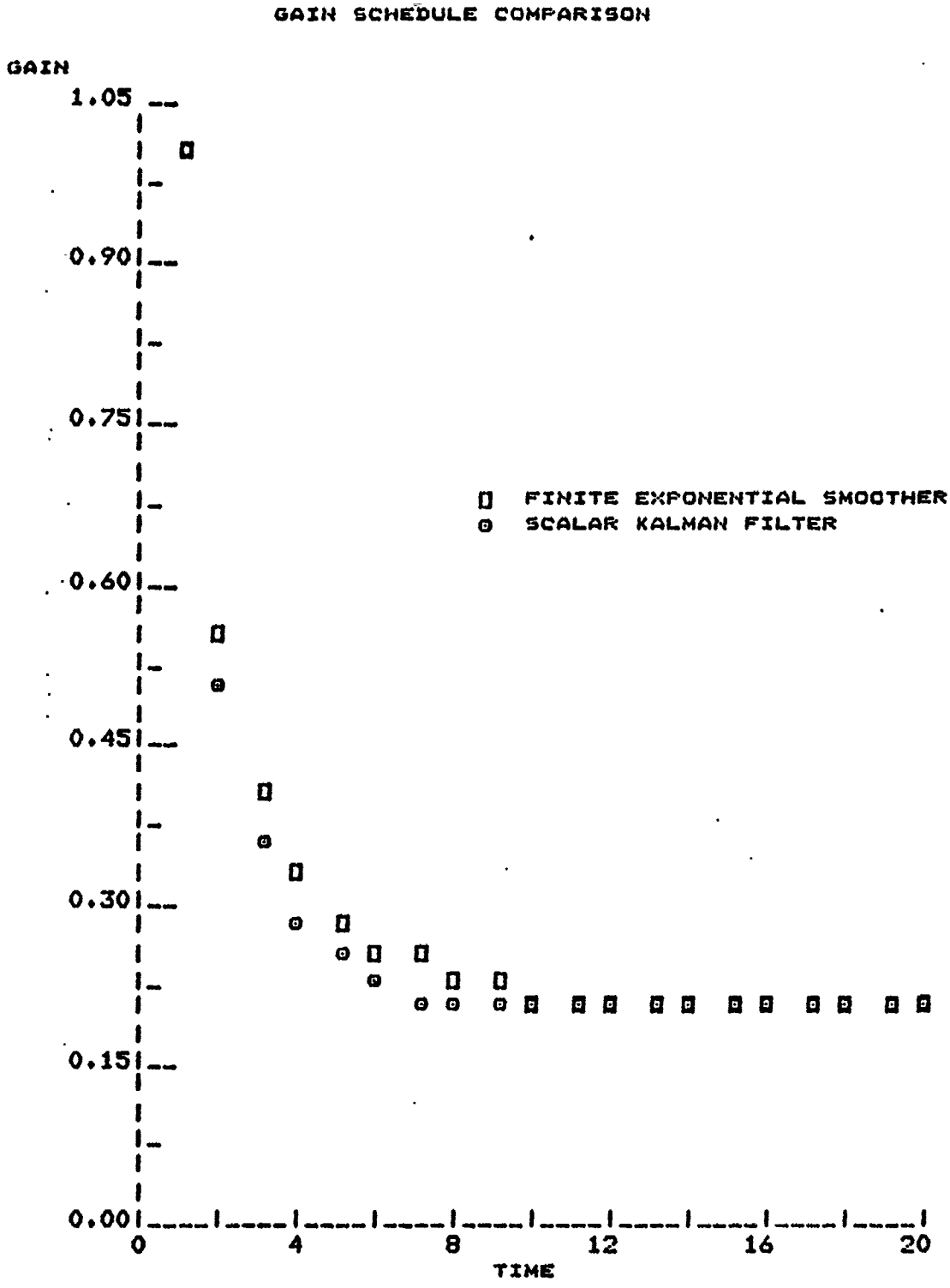

GAIN SCHEDULE COMPARISON
GAIN
1.05
0.90
0.75
0.60
0.45
0.30
0.15
0.00
0
4
8
12
16
20
TIME
FINITE EXPONENTIAL SMOOTHER
SCALAR KALMAN FILTER

FIGURE 13

Kalman filter allows the initial gain to be chosen as less than one, and its value will be determined by the assumed covariance of the prior estimate. Also, the Kalman model forces the analyst to at least think about the concepts of measurement noise and process noise, and to estimate the noise variance ratio Q/R.

As noted before, a frequency analysis of the transient case is not appropriate. However, it can be thought of as a case of transient bandwidth. The filter is initially set to a higher gain than steady-state gain. If there is no prior estimate available, the initial gain is one, and the filter is initially an infinite-bandwidth or all-pass filter. As data are acquired gain drops and the bandwidth narrows until steady-state conditions are achieved. The concept of transient bandwidth is important to the subject of adaptive filtering, to which we will return.

F. HIGHER ORDER FILTERS

The main beauty of the Kalman filter is not in its statistically unbiased method of calculating gain, but in its powerful matrix formulation, which allows it to be applied as a multi-dimensional model incorporating any order of differencing desired. As the state space is increased, it quickly becomes impossible to analyze the filter analytically. High-order multi-dimension filters can also easily exceed the capacity of present digital computers for real-time applications. Fortunately, it has been found that

the state space can be reduced and dimensions decoupled with very little degradation in the overall accuracy of the state estimate [refs. 4 and 5]. For example, if a 12-state model can be reduced to 9 states and can be adequately represented by three 3-state models, the matrix calculations can be considerably simplified and speeded up.

We will examine a second-order (first difference) filter, which can be used to estimate trend, or velocity. We will use the latter term. Position and velocity are to be estimated based only on successive measurements of position. The state transition and the observation matrices are

$$\Phi = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$H = [1 \ 0]$$

The covariance matrices Σ, P, and Q are, of course, 2 by 2 matrices. The state vector has two elements, velocity and position, while the measurement vector has only position. The measurement error R is a 1 by 1 matrix which we will vary. We have chosen Q as

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

arguing that any process noise will be contained entirely in velocity. That is, there can be no random motion that is not caused by a random velocity. Randomness of velocity will feed into position through the state transition matrix.

Even in this simple case, solving analytically for steady-state gain in terms of R and Q requires solving a system of 4th order polynomial equations. We will opt instead for a computer solution. The reader may continue to think in terms of the noise variance ratio, where R will take on the values 1, 10, 100 and Q will remain constant as above. Since there is only one non-zero term in the Q matrix, we may think of the noise variance ratio as the scalar quantity Q(2,2)/R. The resulting steady-state gains are

Noise Variance Ratio	Position Gain	Velocity Gain
1.0	.769	.481
0.1	.553	.211
0.01	.362	.080

As would be expected, the position gain is much higher than that of a scalar filter at an equivalent noise variance ratio, because the process variation now applies to velocity rather than position. The velocity gain is considerably less than the position gain, since the velocity is not measured directly but must be estimated from successive measurements of position. The impulse-response function of the medium-gain filter (noise variance ratio 0.1) is presented in figure 14.

The amplitude response of these three filters is compared in figures 15 and 16. The most striking feature is the amplification which occurs at a specific frequency in the position frequency response. This implies that the filter is

VELOCITY FILTER IMPULSE-RESPONSE

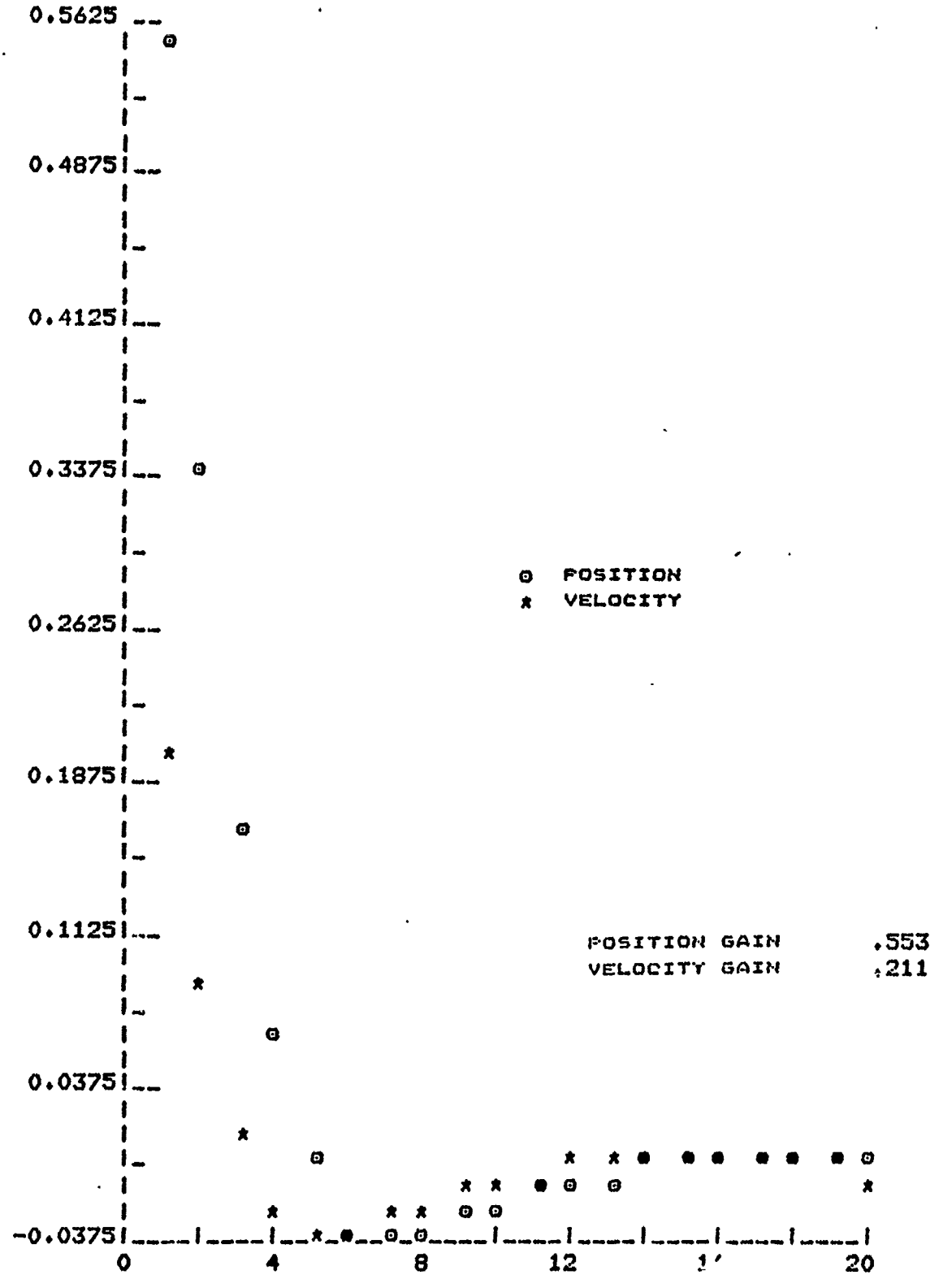


FIGURE 14

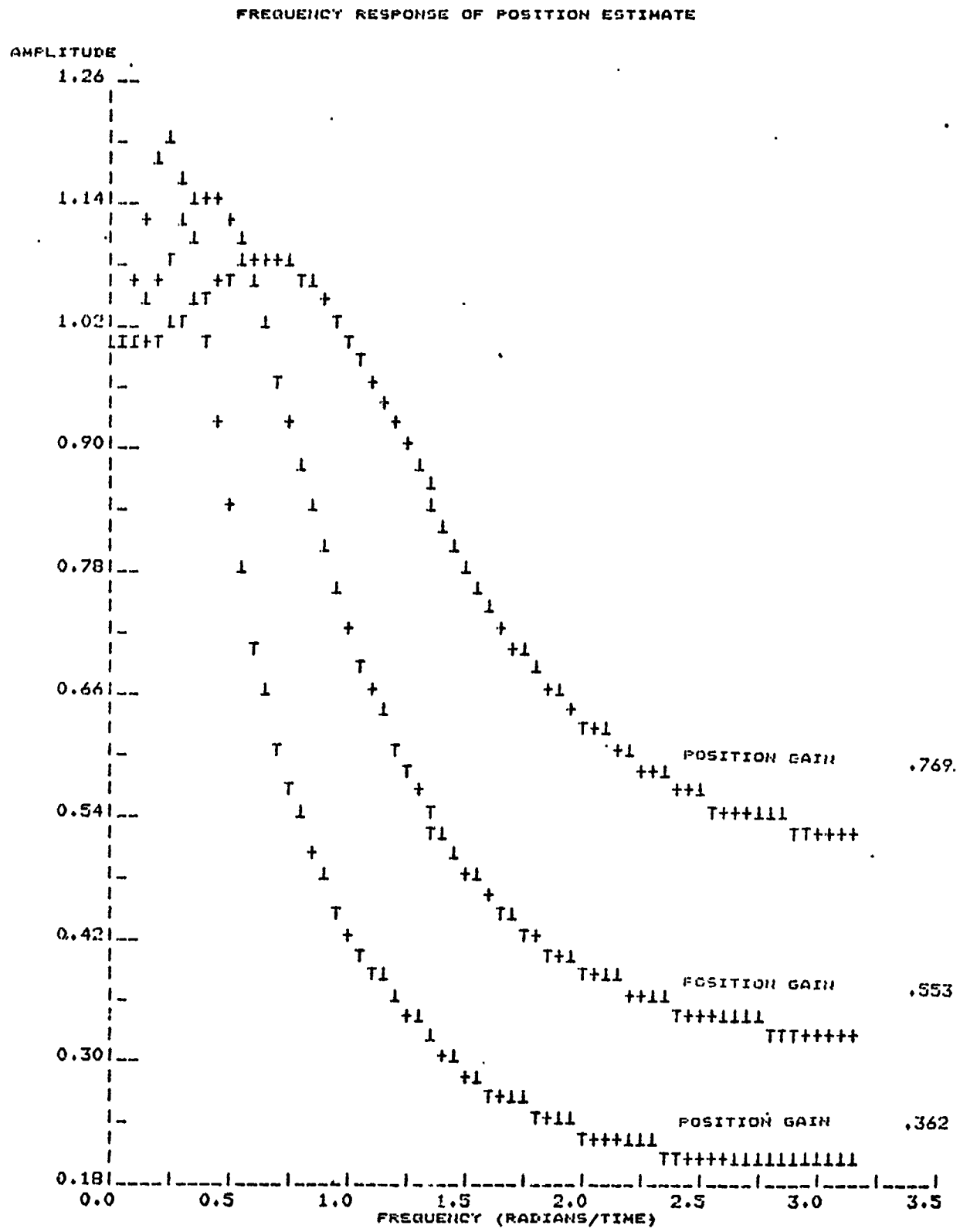

FREQUENCY RESPONSE OF POSITION ESTIMATE
AMPLITUDE
1.26
1.14
1.02
0.90
0.78
0.66
0.54
0.42
0.30
0.18
POSITION GAIN .769.
POSITION GAIN .553
POSITION GAIN .362
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 15

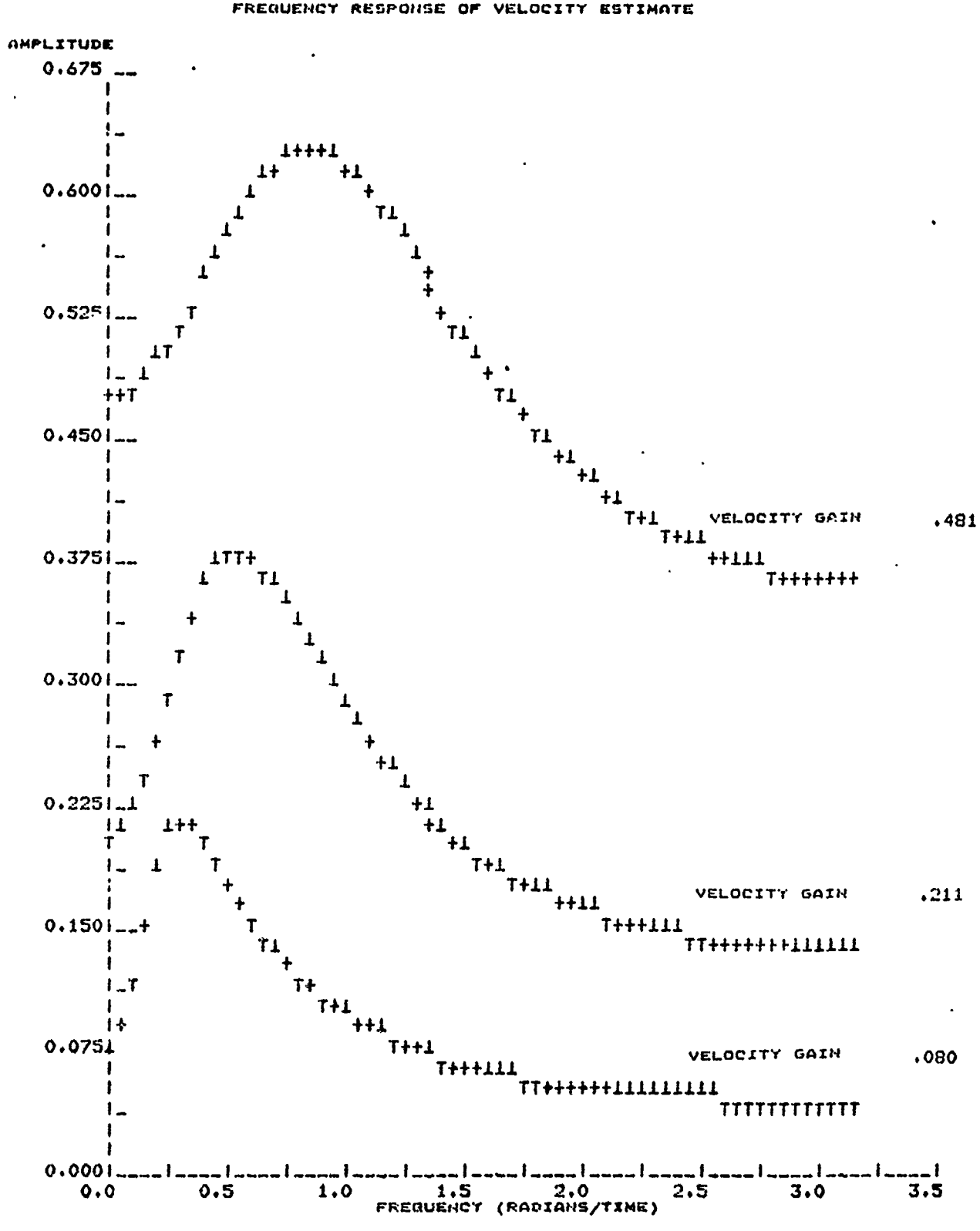

FREQUENCY RESPONSE OF VELOCITY ESTIMATE
AMPLITUDE
0.675
0.600
0.525
0.450
0.375
0.300
0.225
0.150
0.075
0.000
VELOCITY GAIN .481
VELOCITY GAIN .211
VELOCITY GAIN .080
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
FREQUENCY (RADIANS/TIME)

FIGURE 16

most sensitive to motion in a particular frequency range. Thus, the natural frequency of the system to be observed, if it is known, is a significant design parameter. Again, we observe the attenuation of high-frequency noise, although a significant amount still remains at the maximum frequency (note that figure 16 does not include the origin).

The frequency response of velocity shows a reduction in amplitude at low frequency. The amplitude at zero frequency is equal to the velocity gain. This is far from ideal performance for a differentiator, which should have an amplitude response of zero at zero frequency, with a slope of one up to the cutoff frequency [ref.7]. The differentiator is, however, reasonably effective at reducing the amplitude of high-frequency components.

The phase shift of the filters again shows increasing phase lag as gain is decreased. The overall effect is similar to the scalar filter, and is otherwise unremarkable. Therefore, plots are not included.

The data of figure 8 was tested on the lowest-gain velocity filter. Note that the spectral content of the data (figure 9) is quite low-frequency, and that the bandwidth of the lowest gain velocity filter is quite wide, and indeed is higher than that of our scalar filter. So it might be expected that the velocity filter would have some trouble with the data.

The velocity filter performance on the nominal function only is presented in figure 17. The filter overshoots quite badly at the discontinuous steps, which, of course, are an

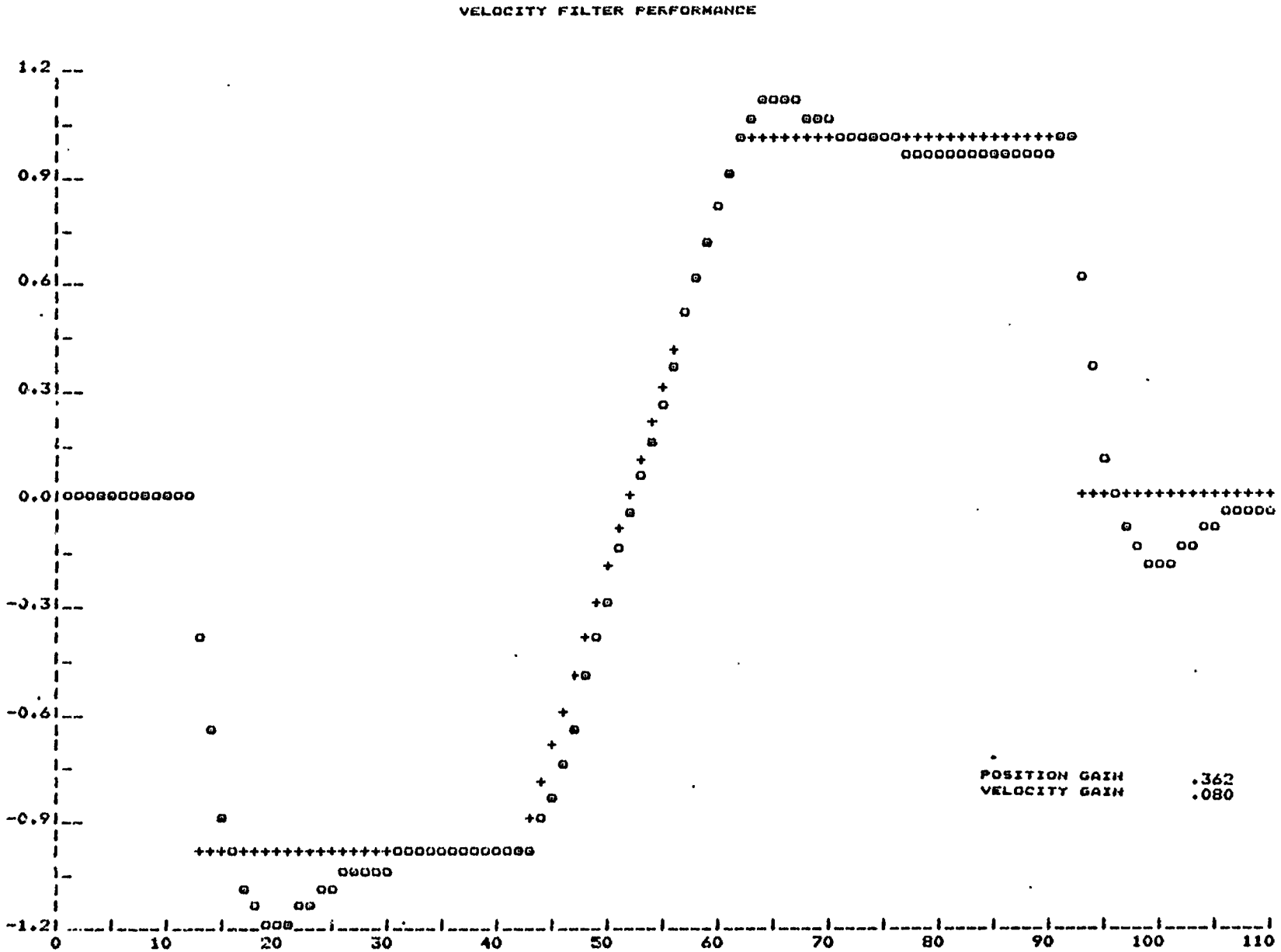


FIGURE 17

SCALE FACTOR FOR ORDINATE: 10

impulse in velocity. The overshoot is less at the start and stop of the trend. It does settle down and track the trend without lag, which is an improvement in performance over the scalar filter. It should be noted that with higher gain, the filter would track the nominal data better, while with lower gain, the overshoots would be more severe.

The performance of the velocity filter on the data is illustrated in figure 18. As expected, the filter tends to follow the noise too much. However, it does follow the discontinuities much more quickly than the scalar filter. This points out the fact that the higher-order filter is more effective as a maneuver detector but it is less suitable for smoothing very noisy data. This again illustrates the concept of bandwidth, which is quite high even in the low-gain velocity filter.

In retrospect, the decision to choose Q(1,1) as zero may not have been wise. Allowing some process noise in position, exclusive of velocity, could well have some smoothing effect on the velocity estimate, which would result in smoother one-period ahead predictions. This could smooth the operation of the filter a bit. The possible combinations of filter parameters, even for this simple filter, are quite numerous.

The frequency response of a second-difference (acceleration) filter was also determined for comparison. The results are presented in figure 19. The Q matrix was zero except for Q(3,3), which was one. R was chosen as 10, resulting in a nominal noise variance ratio of 0.1. Again,

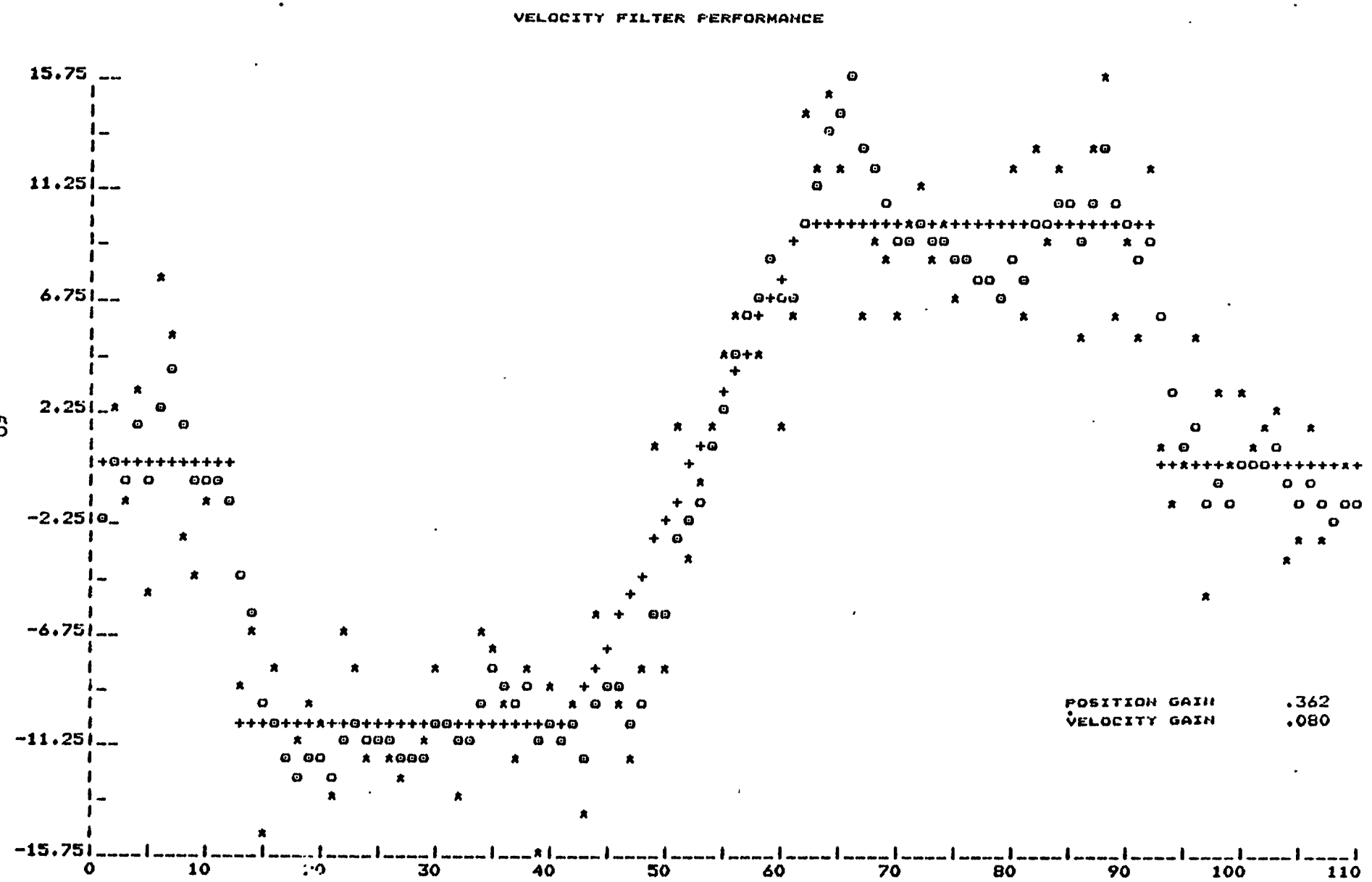

VELOCITY FILTER PERFORMANCE
15.75
11.25
6.75
2.25
-2.25
-6.75
-11.25
-15.75
0
10
20
30
40
50
60
70
80
90
100
110
POSITION GAIN .362
VELOCITY GAIN .080

FIGURE 18

the gain was higher than that of a velocity filter with an equivalent noise variance ratio. The amplification of low-frequency components of position was increased, and the zero-frequency amplitudes of velocity and acceleration again corresponded to filter gain.

FREQUENCY RESPONSE OF ACCELERATION FILTER

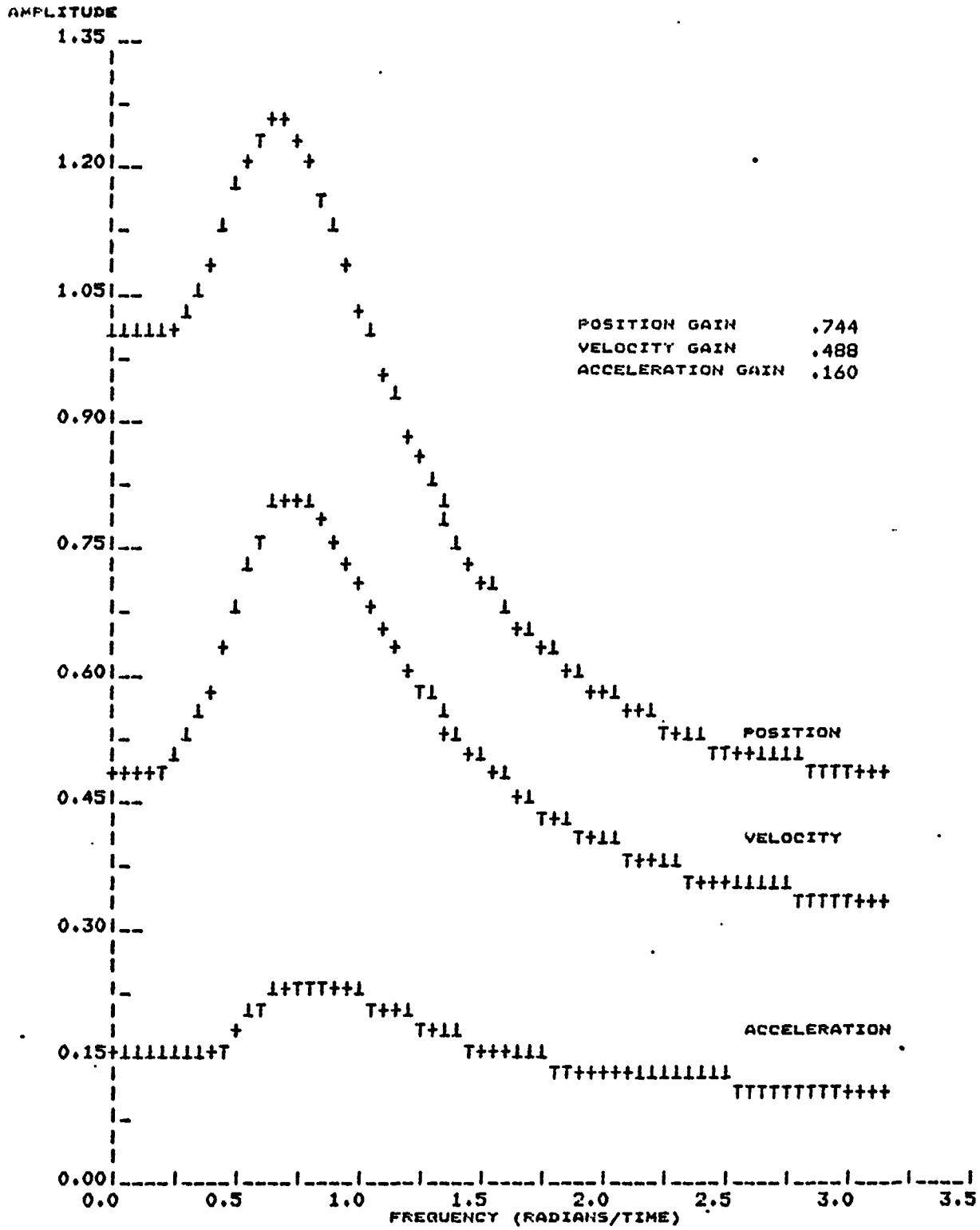


FIGURE 19

IV. ESTIMATION, SMOOTHING, AND PREDICTION

A. ESTIMATION

So far, we have been concerned only with estimation of the present state. A filter designed to provide such an estimate cannot be symmetric, because it can put no weight on future observations. Thus, phase lag is inevitable, and is one of the parameters that should be considered in the design process.

B. SMOOTHING

Smoothing is the use of a filter to provide an estimate of past states. Such a filter can be made symmetric, which completely eliminates the phase lag. Non-recursive smoothing filters cause a loss of N data points at each end of the data, where the span of the filter is (-N, N).

The Kalman filter can be used as a smoother by simply running the forward estimate through the filter in the opposite direction. The impulse response function of the scalar filter was

$$g(t) = ab^{t}$$

It can be shown (appendix A) that the impulse response

function of the smoother (forward and backward filters combined) is

$$G(t) = ab^{t}/(1+b)$$

which is just the convolution

$$g(t) \circledast g(-t)$$

The Kalman filter is able to provide an estimate throughout the span of the data. No data is lost at either end. However, due to transient effects, the data near either end is subject to phase shift and some increase in gain. The filter is necessarily not symmetric near each end of the data span.

Gelb [ref.4] includes a complete discussion of fixed-point, fixed-lag, and fixed-interval smoothing. We will restrict our attention to the scalar, fixed-interval, steady-state case, ignoring the end effects.

The scalar Kalman filter of section III.D (noise variance ratio of 0.1, gain of 0.27) was used as a smoother on the data of figure 8. The results are presented in figure 20. As compared to the one-pass performance as illustrated in figure 12, the smoothed data shows phase lag removed and peaks in the oscillations reduced. However, the smoother has less ability to follow the discontinuities in the nominal function. The removal of the phase lag is characteristic of any symmetric filter. However, the reduced

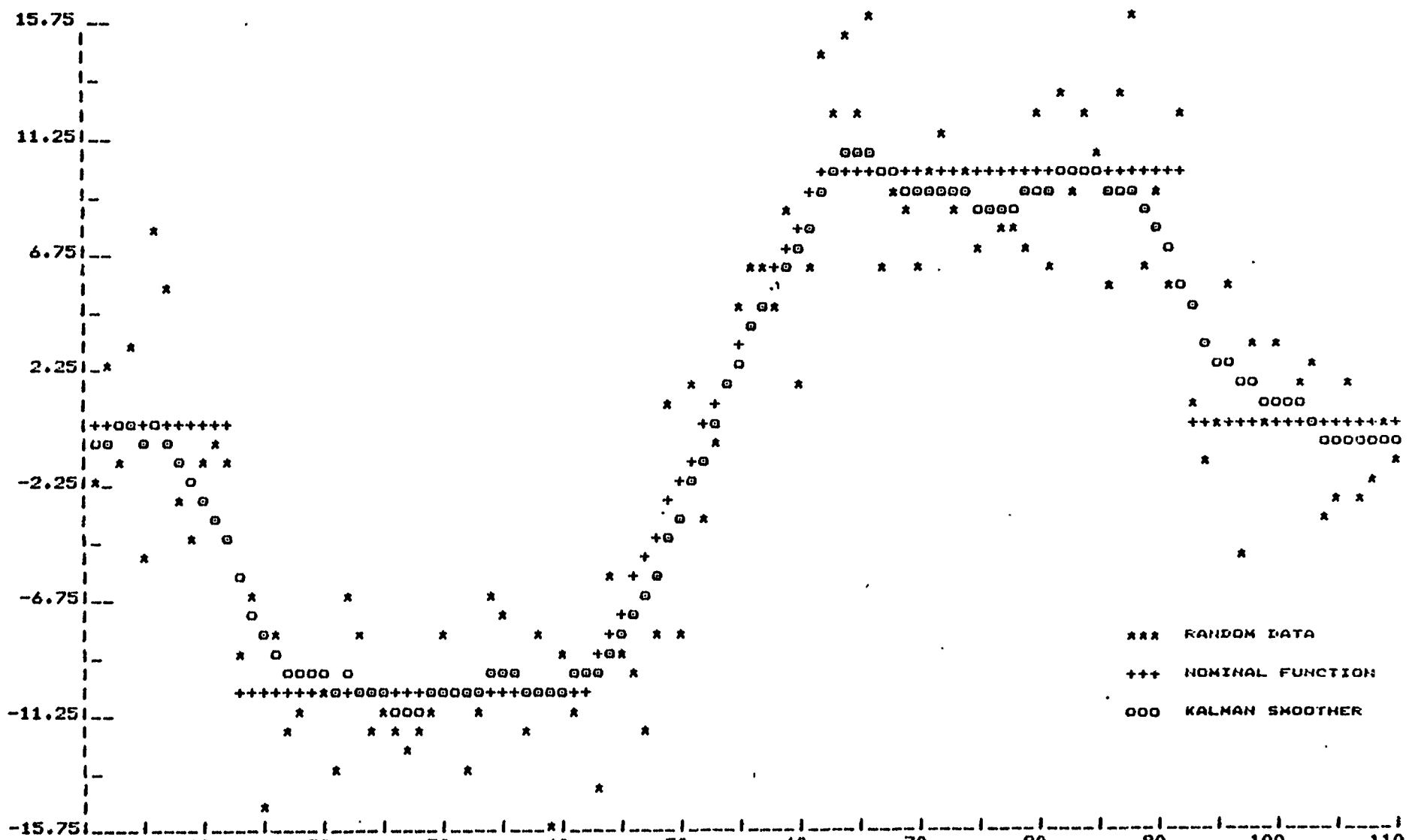

KALMAN SMOOTHER PERFORMANCE
15.75
11.25
6.75
2.25
-2.25
-6.75
-11.25
-15.75
0
10
20
30
40
50
60
70
80
90
100
110
*** RANDOM DATA
+++ NOMINAL FUNCTION
ooo KALMAN SMOOTHER

FIGURE 20

ability to follow the discontinuities is the result of reduced effective gain. Since we have convolved the filter weights, we have squared the amplitude of the frequency response. The weight on the data point at (t=0) is reduced. The effective gain was 0.27 for the forward filter, and 0.156 for the smoothing filter. This reduction in effective gain is not addressed in the literature on the Kalman filter, and it is unclear how smoother gain should be chosen in relation to the noise variance ratio Q/R.

C. A COMPARISON OF TWO SMOOTHERS

The Kalman smoother of the previous section was compared with a Gaussian smoother to illustrate some design options and procedures. The Gaussian smoother was chosen from among a huge variety of data windows because it has good smoothing properties, and because it is particularly easy to design. Interestingly, preliminary experiments showed repetitive applications of a Kalman filter to result in an approximately Gaussian filter weight distribution. A comparison of the Gaussian smoother to a variety of other windows is contained in Harris [ref.15].

The Gaussian smoother is a symmetric filter with the weights chosen according to a discretized and truncated normal distribution. The formula is

$$g(t) = K \exp(-t^2/2\sigma^2)$$

where t is an integer on the range (-N, N) and K is chosen such that

$$\sum g(t) = 1$$

The ease of design comes from the observation that the Fourier transform of the continuous Normal distribution is also a Normal distribution with scale parameter (variance) equal to $1/\sigma^2$. As long as ($\sigma > 2$) and truncation is not more severe than $|N| > 2\sigma$, a reasonable approximation of the frequency response for the Gaussian digital filter is

$$G(v) \approx \exp(-\sigma^2 v^2/2)$$

The scale parameter was chosen such that the frequency response was equal at ($v = 0.5$). Skipping the algebraic details, this required ($\sigma = 3.11$). The Gaussian smoother was truncated to ($N = 7$), resulting in a filter span of 15 data points. The frequency response of both filters is presented in figure 21, and the filter weighting coefficients are presented in figure 22.

Since we truncated the Gaussian smoother, we would expect some ripples in the tail of the frequency response, which are just barely visible in figure 21. The Gaussian filter has a sharper transition band, and is quite effective in blocking high frequencies. As compared to the Kalman smoother, the Gaussian filter weights the present data point

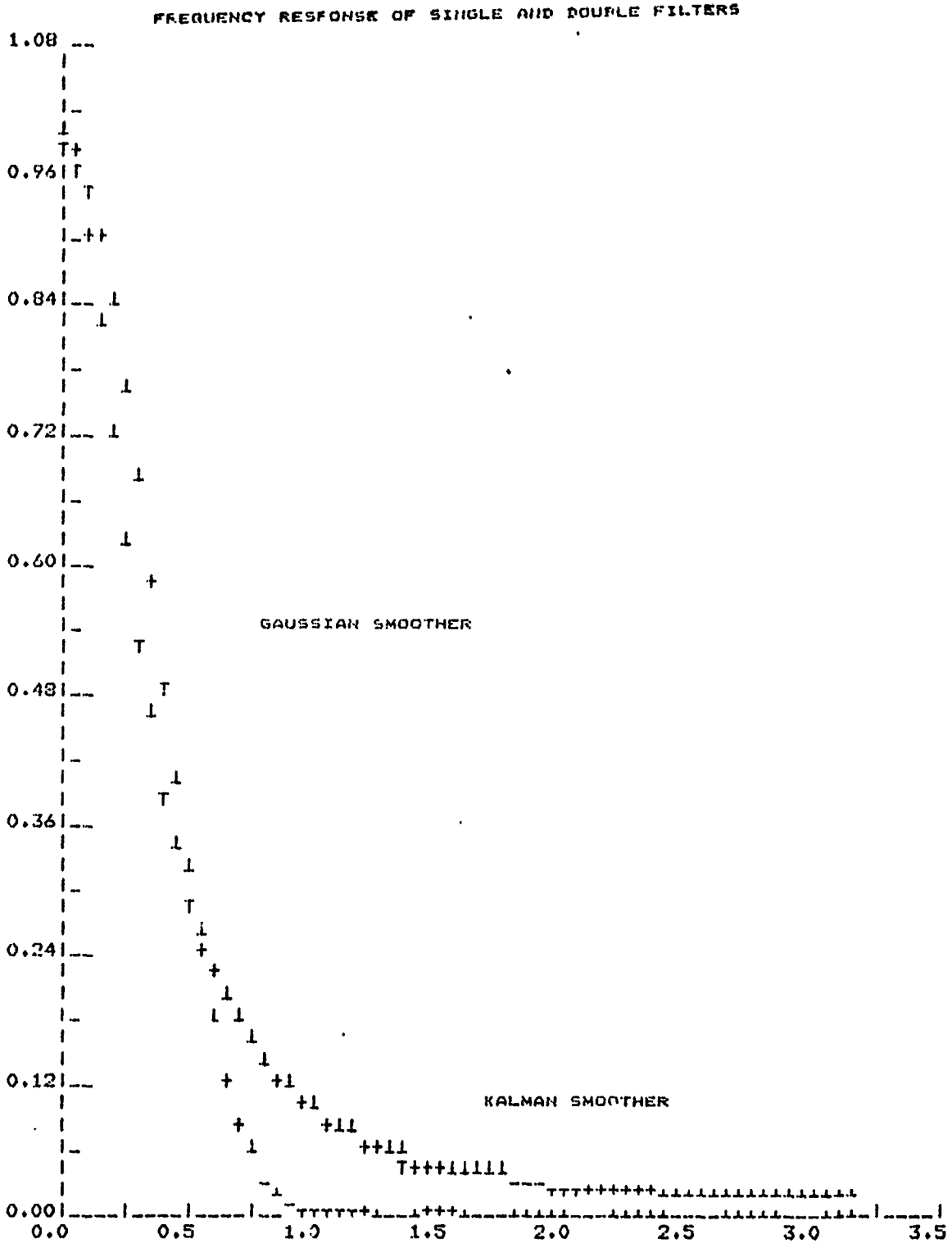

FREQUENCY RESPONSE OF SINGLE AND DOUBLE FILTERS
1.08
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
GAUSSIAN SMOOTHER
KALMAN SMOOTHER
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

FIGURE 21

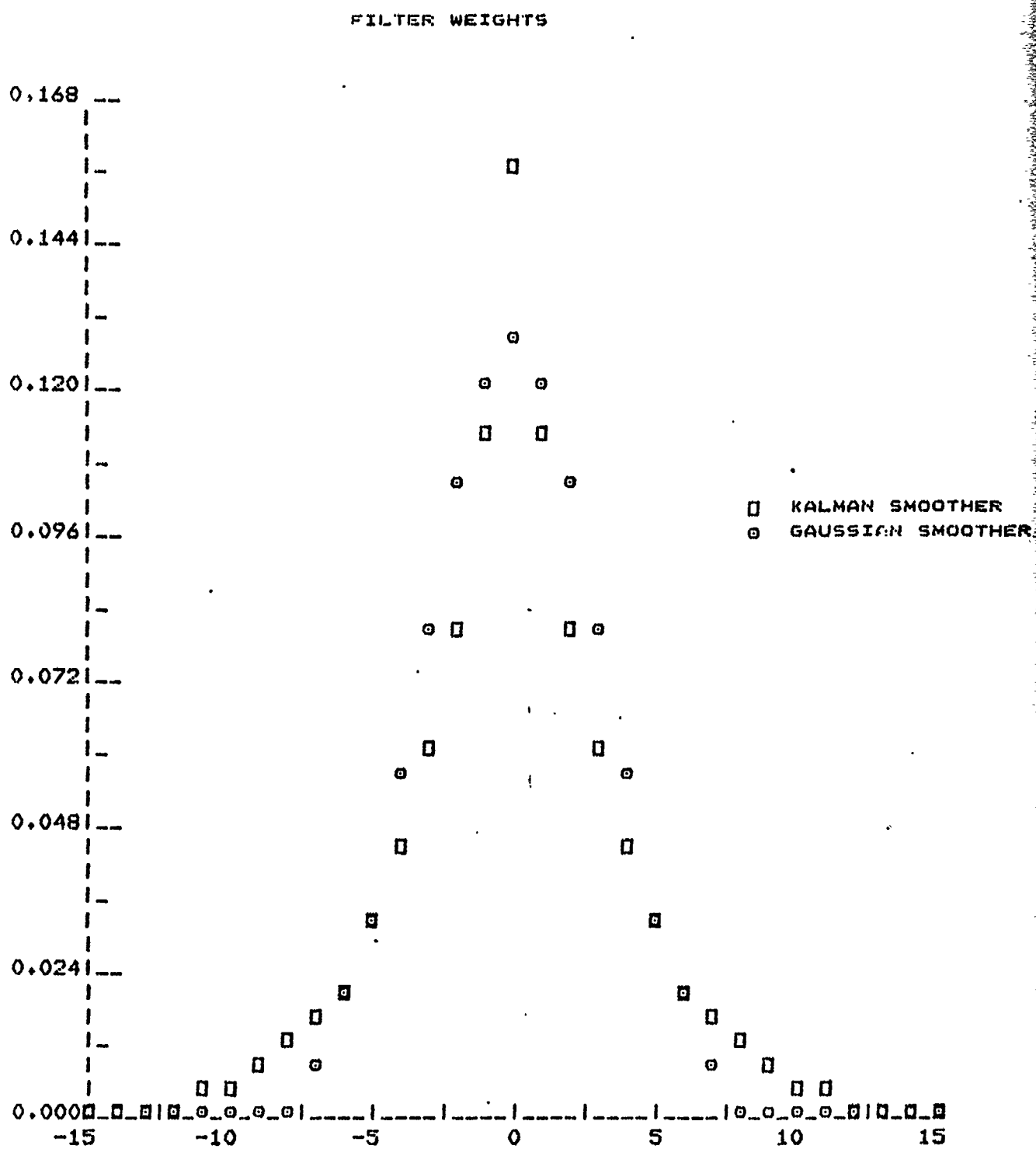

FILTER WEIGHTS
0.168
0.144
0.120
0.096
0.072
0.048
0.024
0.000
-15
-10
-5
0
5
10
15
KALMAN SMOOTHER
GAUSSIAN SMOOTHER

FIGURE 22

less, nearby data points more and farther data points less.

The performance of the Gaussian smoother on the data of figure 8 is presented in figure 23. A comparison of the Kalman and Gaussian smoothers is presented in figures 24 and 25. In figure 24, the smoothers are applied only to the nominal function. It can be seen that the Gaussian smoother followed the discontinuities and corners of the nominal function better than the Kalman smoother. However, when the smoothers were applied to the noisy data, the results were less clear (figure 25). The Gaussian smoother again followed the nominal function a bit better, but it also followed low-frequency components of the noise a bit more, tending to emphasize cyclic effects that aren't really there. The mean square difference between the smoothed estimate and the nominal function were very similar, 1.81 for the Gaussian smoother and 1.84 for the Kalman smoother. Thus, the Kalman filter seems quite effective when used as a smoother. The reader is reminded that the mean square difference for the scalar Kalman filter was 5.1, which clearly indicates the superiority of smoothing over filtering.

D. PREDICTION

Prediction is difficult. Recall that a stochastic process is an ensemble of possible paths, while data is the manifestation of one member of that ensemble. What could have happened did, but what can happen isn't necessarily going to. Prediction can be thought of as filtering without

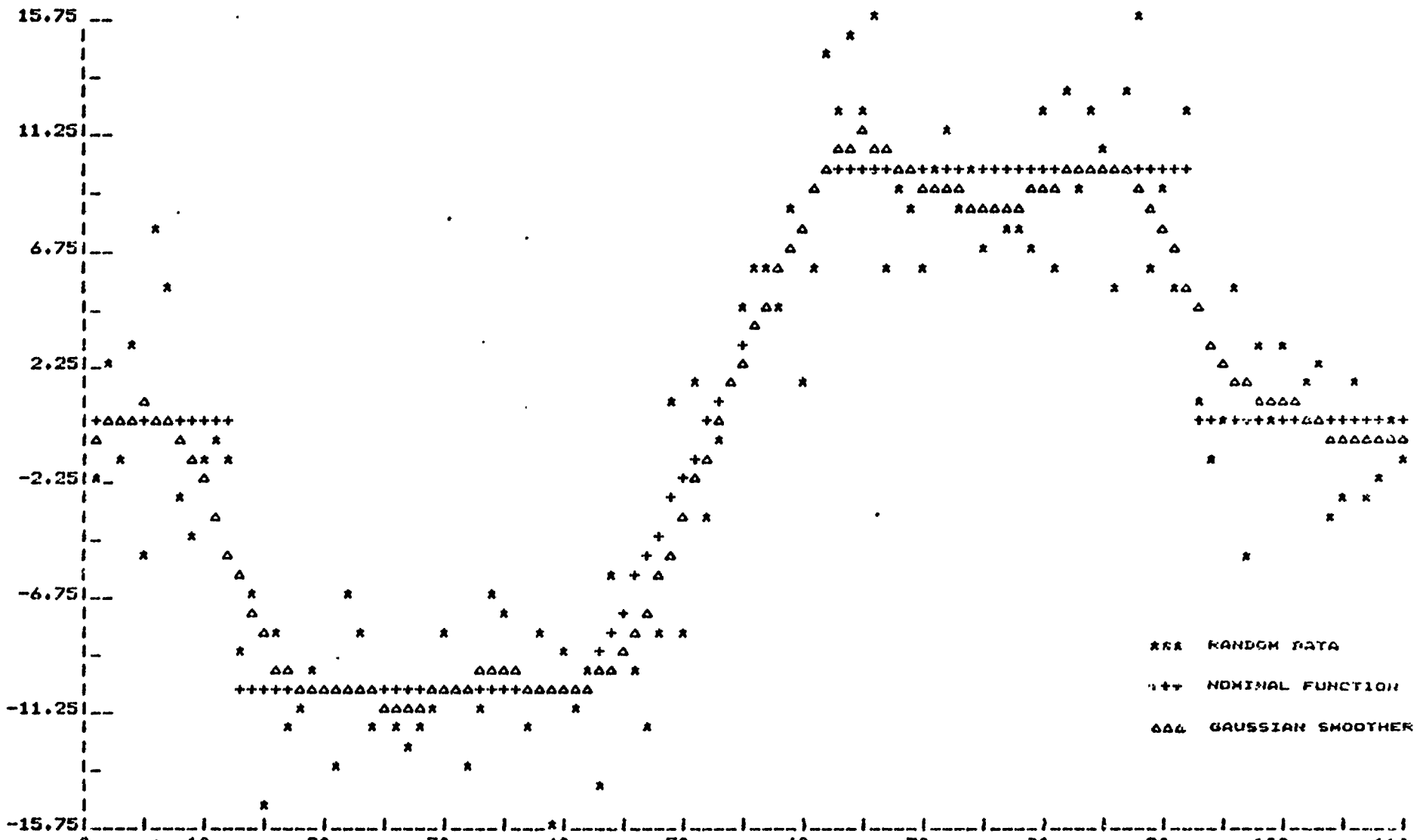

GAUSSIAN SMOOTHER PERFORMANCE
15.75
11.25
6.75
2.25
-2.25
-6.75
-11.25
-15.75
0
10
20
30
40
50
60
70
80
90
100
110
*** RANDOM DATA
+++ NOMINAL FUNCTION
ΔΔΔ GAUSSIAN SMOOTHER

FIGURE 23

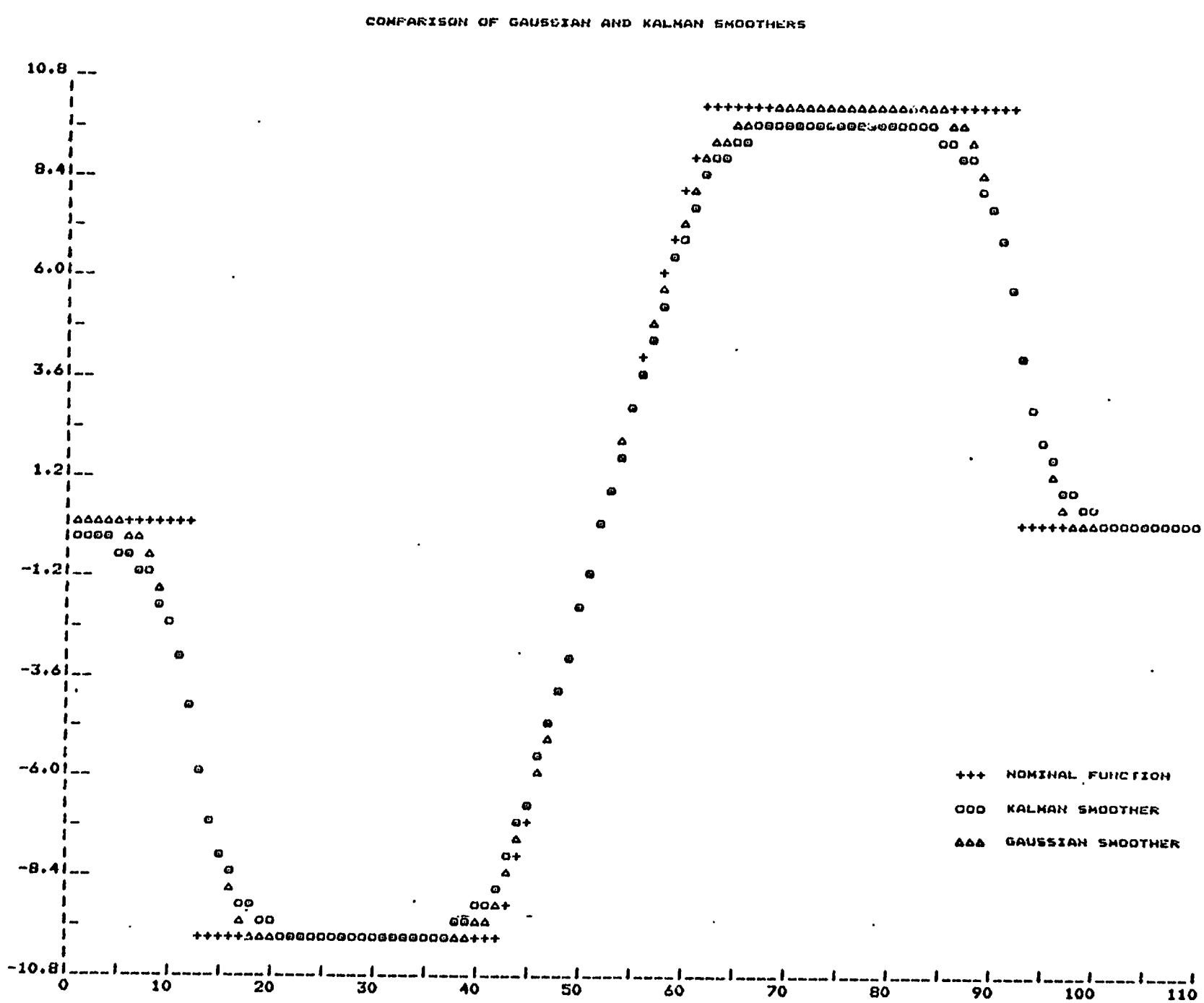

COMPARISON OF GAUSSIAN AND KALMAN SMOOTHERS
10.8
8.4
6.0
3.6
1.2
-1.2
-3.6
-6.0
-8.4
-10.8
0
10
20
30
40
50
60
70
80
90
100
110
+++ NOMINAL FUNCTION
OOO KALMAN SMOOTHER
ΔΔΔ GAUSSIAN SMOOTHER

FIGURE 24

COMPARISON OF GAUSSIAN AND KALMAN SMOOTHERS

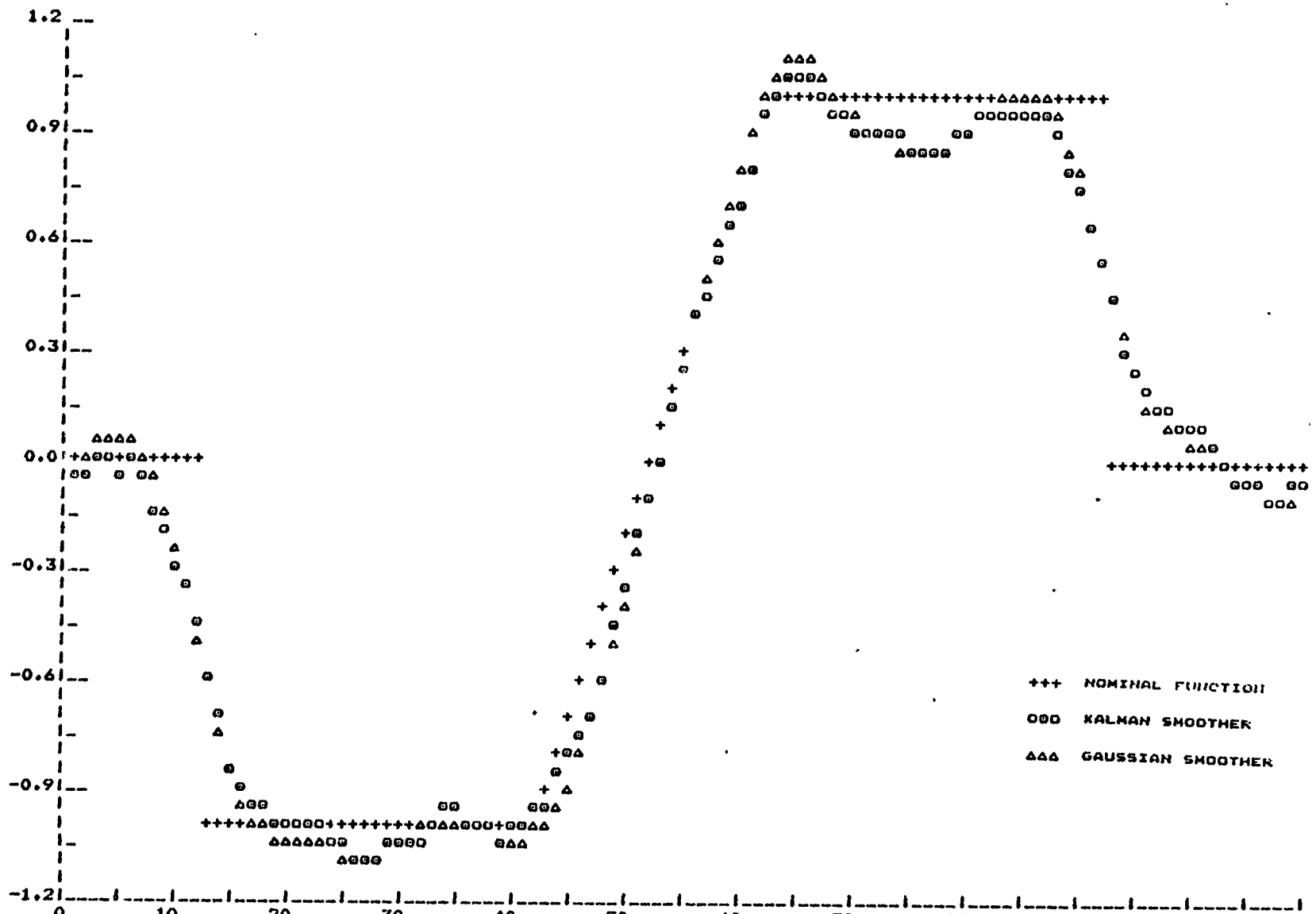


FIGURE 25

SCALE FACTOR FOR ORDINATE: 10

measurements, and the estimate is projected forward in time through the state transition model. Filtered or smoothed estimates may be quite accurate, even if the state transition model is not. However, good predictions are heavily dependent on an accurate state transition model.

The higher-order Kalman filters are polynomial models. Hamming has pointed out that polynomial models are poor predictors, since the estimate tends to veer off to plus or minus infinity as soon as the model is released from the data [ref.7]. There is no requirement to use the same model for prediction as for filtering. For example, it might make sense to track a target with an acceleration filter, but to compute fire control information based on a constant-velocity model, since target acceleration is generally assumed random with zero mean. Similarly, the economist may desire to filter data with a high-order model, but make predictions based on constant trend. Clark [ref.5] discusses a somewhat more sophisticated method due to Singer, in which the model decays from an acceleration predictor to a constant-velocity predictor as prediction time increases. Such techniques are heuristic in nature, but can prove valuable to the innovative analyst.

An interesting example of the above concept can be found in Box and Jenkins [ref.11]. They compared a quadratic forecast due to Brown [ref.10] with their own IMA(0,1,1) model with a gain of 0.9. The latter model is equivalent to the steady-state scalar Kalman filter. The data used for the comparison was a time series of IBM stock prices. Box and

Jenkins observed that, while the quadratic model might well be used to fit the data, its performance as a predictor was clearly inferior to the simpler IMA(0,1,1) model. This is not surprising, since it has long been suggested that stock prices behave as a random walk, and that the best forecast of stock price, at least in the short run, is the present price [ref.11]. Note that the foregoing implies that the gain should be set to 1.0, which corresponds to no filtering at all. Therefore, Box and Jenkins apparently found that some filtering of the data was appropriate, even though the gain they used was quite high.

V. SOME REFINEMENTS, EXTENSIONS AND ALTERNATIVES

A. ADAPTIVE FILTERING

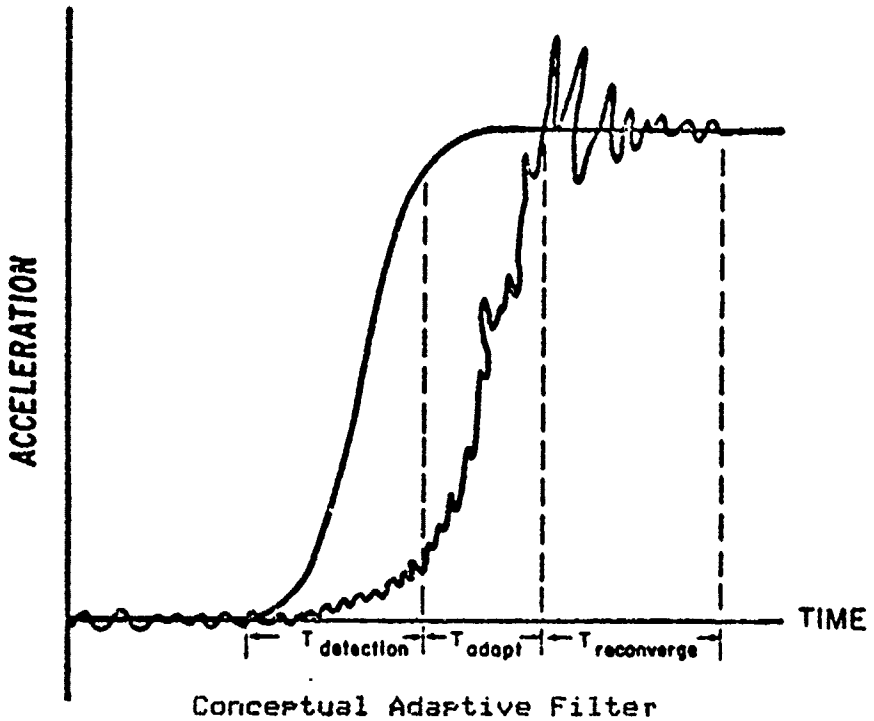
In the linear Kalman filter, the gain is completely independent of the data. Clearly, this will result in major errors if the gain is chosen inappropriately or if the data statistics change. If the gain is too low, the filter lags badly. In the extreme case, which occurs if the process noise covariance Q is much too low, the filter pays much too much attention to the past and diverges from the data. On the other hand, if the gain is too high, the filter pays too much attention to the data and the state estimate contains noise. If the filter is a polynomial model and is to be used as a predictor, the resulting errors will be spectacular.

The solution is simple in concept but can be difficult to implement. One simply sets the steady-state gain as low as appropriate for the stable process being estimated. In target tracking, the gain would be set to track an airplane flying a straight path. A "maneuver detector" or "trend detector" is incorporated, which is nothing more than a recursive statistical test applied to the residuals to determine whether or not they come from a zero-mean distribution. If not, the bandwidth is gradually widened (gain is increased) until the residuals pass the zero-mean test. Then, the gain is allowed to decrease toward the stable, steady-state value.

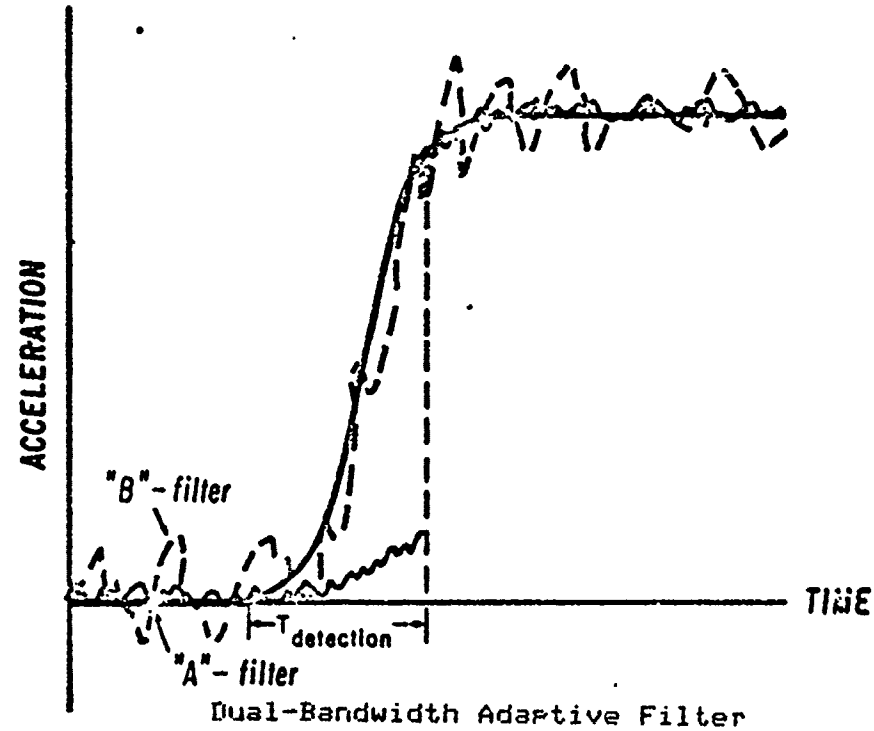
Further details and some novel approaches are discussed in Clark [ref.5]. Two examples taken from Clark are illustrated in figure 26. The conceptual adaptive filter discussed above requires time to detect the maneuver or trend, adapt to it, and reconverge to a stable gain setting. During this time, the state estimate is less accurate, and the time required may be unacceptably long for some applications.

Clark proposes a dual-bandwidth adaptive filter to speed adaptation. The process is simultaneously tracked by a narrow-band and a wide-band filter. If a maneuver or trend is detected, the state estimate of the wide-band filter is fed into the narrow-band filter. Ideally, this would allow the narrow-band filter to jump immediately to the current (unbiased) estimate of the wide-band filter. In practice, Clark found that some widening of the bandwidth of the narrow-band filter was also required.

Voluminous literature exists on the subject, much of it very difficult to read. Clark [ref.5] incorporates a particularly lucid account of stability problems encountered, methods of reducing the cost of false detection of bias, analytical methods of determining filter parameters, and experimental results. Although Clark's filter was designed to track and predict the position of airborne targets, the methods discussed are adaptable to the filtering of economic time series or virtually any other stochastic process.



Conceptual Adaptive Filter



Dual-Bandwidth Adaptive Filter

Figure 26. Adaptive Filters

B. NON-LINEAR FILTERING

In the non-linear Kalman filter, one or more of the matrices Q, R, H, or Φ are allowed to vary with time. Since this results in a time-variation of the gain matrix, quantitative analysis in the frequency domain is no longer appropriate. However, it is well to keep the concepts in mind in order to gain additional insight. There are two basic types of non-linearities that may arise; non-linear measurements and non-linear dynamics.

1. Non-Linear Measurements

Non-linear measurements arise when observations are made in one coordinate system and the model requires that the state be estimated in another coordinate system. In this case, the matrices R and H are time-varying functions of the coordinate transformation, and do depend on the data, in the sense that they depend on the location of the data within the coordinate system. This type of non-linearity is often easy to handle.

For the best example of non-linear measurements we must return to the target-tracking model. Fire control systems generally track in azimuth, elevation, and range. However, the model is a polynomial in Cartesian coordinates, but not in polar coordinates. Airplanes often fly a straight path, but seldom, if ever, fly a constant bearing or range with respect to the radar observer. In this model, the

non-linearity can be reduced by considering the Cartesian measurement error as a linear transformation of the polar measurement error. If the polar measurement error is Gaussian, the Cartesian measurement error is very nearly Gaussian with covariance matrix R a function of the coordinate transformation.

The Cartesian R matrix will not be diagonal, even if the polar R matrix is. However, Clark [ref.5] has found that setting the off-diagonal terms of the Cartesian R matrix to zero did not appreciably degrade filter performance. In this way, he was able to decouple a nine-state filter into three three-state filters.

If the measurement non-linearity is too severe, it may not be reasonable to assume that the noise is Gaussian. However, limited experiments performed on data with non-Gaussian noise (an exponential distribution was used) showed that the Kalman smoother and the Gaussian smoother were quite robust as long as the gain was not high. This seems to be a consequence of the Central Limit Theorem, since low gain implies a linear combination of a fairly large number of data points. It should be noted that a filter designed to handle this situation is still linear, although the Gaussian assumption is violated.

2. Non-Linear Dynamics

Non-linear dynamics are considerably harder to handle than non-linear measurements. This is unfortunate, since the areas of potential application are numerous. Non-linear

dynamics occur when the Q or Φ matrices depend on the previous history of the process.

As a simple example, consider the multiple regression model

$$y(t) = a\, x_1(t) + b\, x_2(t) + y(t-1)$$

where $y(t-1)$ corresponds to the intercept term. In this model, we wish to estimate the dependent variable $y(t)$. To do so, we need to estimate not only the independent variables $x_1(t)$ and $x_2(t)$, but also the regression coefficients a and b. Let us assume we can measure $y(t)$, $x_1(t)$, and $x_2(t)$, but not a and b. Assuming a first-order system, the state update equation is

$$\begin{bmatrix} y(t+1|t) \\ x_1(t+1|t) \\ x_2(t+1|t) \\ a(t+1|t) \\ b(t+1|t) \end{bmatrix} = \begin{bmatrix} 1 & a(t|t) & b(t|t) & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y(t|t) \\ x_1(t|t) \\ x_2(t|t) \\ a(t|t) \\ b(t|t) \end{bmatrix}$$

Where the transition matrix is unfortunately not unique. The first row of the transition matrix could be equally well represented by

$$[\, 1 \quad 0 \quad 0 \quad x_1(t|t) \quad x_2(t|t) \,]$$

or even

$$[\ 1 \quad a(t|t)/2 \quad b(t|t)/2 \quad x_1(t|t)/2 \quad x_2(t|t)/2\ ]$$

and it obviously changes at every iteration. It is at this point that filter design becomes an art.

Note that the independent variables and the regression coefficients are assumed here to be first-order Gauss-Markov processes. Increasingly high orders would multiply the state space.

Several experiments were run using a second-order model similar to the above on the Box-Jenkins [ref.11] series M data (sales data with leading indicator). Quantitative results are not presented, because the Box-Jenkins data did not include sufficient forecast estimates for comparison, some "cheating" was done because the Box-Jenkins parameters were used in filter design, and it never became clear exactly what parameters were appropriate for the R and Q matrices. However, some qualitative comments are appropriate. The model did work. Some instability was noted in the regression parameters. It became obvious that the gain on the regression parameters must be set very low in comparison to the gain on the independent variables, in order to keep the regression parameter estimates from varying faster than the estimates of the independent variables. This implies choosing small values for the noise variance of the regression parameters. Also, by keeping the gain fairly low

on the leading indicator, it was possible to induce a phase lag that approximately cancelled the lead.

The requirement to keep gain low in order to improve stability is evidently a consequence of the increased degrees of freedom. The more parameters to be estimated, the more degrees of freedom in the model. High gain is analagous to relatively few data points being used in regression. The more variables we introduce into the model, the less gain we are able to use.

It is indeed unfortunate that multiple regression is a non-linear problem when cast in a filter model. It would be useful to have a multiple regression model for which more recent observations were weighted more heavily than older ones in determining the regression parameters. No doubt the innovative analyst could develop one to fit the specific situation. However, clearly-defined techniques with demonstrated results are not available to the practitioner. The experts all have their favorite methods, and much of the literature is difficult to read. There is clearly a need for additional research in this area.

C. NON-PARAMETRIC FILTERING

We close our discussion with an interesting alternative to conventional digital filtering techniques. There are those who are bothered by the usual distributional assumptions made in any application of parametric statistics. An extensive literature has developed in the field of

non-parametric statistics, which is based on the principle that distributional assumptions are avoided, or at least weakened. A strong point of non-parametric statistics is relative insensitivity to extreme outliers. However, little progress has been made in the non-parametric analysis of time series. An exception may be found in Tukey [ref.16], which is presented in a highly intuitive manner, with little or no theoretical background.

One simple idea advanced by Tukey is that of median smoothing. The smoothed estimate is based on the median of several adjacent data points, rather than on a weighted linear combination. The result is obviously a series of steps, since adjacent data points will often have the same median. Tukey suggests several methods to restore some curvature in the estimate. These will not be developed here. Tukey's methods would be relatively hard to mechanize on a computer, because the methodology requires extensive logical rules.

Tukey's methods could be extended to real-time filtering problems by developing a non-parametric analog to the recursive digital filter. Recall that the recursive digital filter consists of a weighted linear combination of recent data points added to a weighted linear combination of recent estimates. The non-parametric filter estimate would simply be the median of several recent data points and several recent states. The idea is intuitively appealing and should be the subject of future research. Discussion here will be limited to some of the more obvious traps that await the

unwary.

If such a filter were to be designed, the impulse-response function would be meaningless, because the median estimate would always be zero if the span of the filter were greater than two. Nevertheless, a median filter does have a frequency response, which in fact is a particularly nasty one.

Consider a seven-point median smoother, where the state estimate at time t is the median of the measurements made at time (t-3) to time (t+3). This is analagous to the rectangular (parametric) window discussed in Hamming [ref.7]. The rectangular window weights all data points within the window equally. The median window obviously does the same. As a result, we would expect the frequency response of the median window to have severe ripples as does that of the rectangular window. We can see intuitively that this is true. Since the span of our example median window is 7, the frequency response of any frequency that is a non-zero integer multiple of 1/7 is obviously zero. The frequency response at zero frequency is one, since the zero frequency implies a constant. The amplitude of the frequency response falls off to the first zero, then rises again. Successive maxima decrease with increasing frequency, but the frequency response is always non-zero except at frequencies that are non-zero integer multiples of the reciprocal of the span of the window. Thus, the non-parametric filter will need to incorporate some (unspecified) device to improve the frequency response.

Two other difficulties are worthy of mention. First, the sampling distribution of the median may have a larger variance than the sampling distribution of the mean. This means that the parametric filter may provide a better estimate than the non-parametric filter if the assumptions on which the parametric filter is based are all reasonable. Second, for the non-parametric filter to be useful, the median must be a [illegible] of interest. If it is assumed that the distribution is symmetric, the median and mean are, of course, equal. [illegible] sampled distribution is skewed, the mean cannot be deduced from the median unless strict parametric assumptions are imposed, which of course, override the justification for the non-parametric filter in the first place. The idea is nevertheless intriguing, and should be explored further.

VI. SOME APPLICATIONS

The Kalman filter has been applied to Operations Research and economic problems with varying degrees of success. McWhorter [ref.17] conducted an empirical study of the Kalman filter in which he compared it to several other methods of time series forecasting. The results were mixed, with no method dominating. The Kalman filter compared more favorably over a short term forecasting horizon than over a long term one. Its performance was, not surprisingly, found to be degraded if the structural model was seriously mis-specified. McWhorter pointed out some of the difficulties encountered in building the model. In an economic context, it is often very difficult to specify the noise covariance matrices R and Q, and even to identify the structure of the state transition matrix Φ. The assumptions made are often sweeping and arbitrary, in contrast to tracking applications where the noise processes and especially the state transition model are relatively well understood.

A. INVENTORY MANAGEMENT

The Kalman filter is directly applicable to inventory management, and if properly designed, should be superior to the finite exponential smoothing model of Bessler and Zehna [ref.14]. Downing, Pike, and Morrison [ref.18] designed a

Kalman filter for the inventory control of nuclear material. The paper is readable, and the filter is well-documented and easy to understand. They use the concept of a control vector, which has not been mentioned here. An interesting peculiarity of the model is that one of the measurements is only available once every twenty iterations. The state transition matrix is a simple material balance relation which is obviously quite accurate. Such a model could be expected to perform quite well.

B. ESTIMATING A MEAN FUNCTION

Although the Kalman filter was derived from an assumption of stationarity, we have seen that it can be quite powerful in separating a time varying signal from noise. The examples of section III were all essentially estimates of the time-varying mean function of a stochastic process. The example process was Gaussian with a constant variance. The variance was the measurement noise, and so directly influenced the gain. If variance were not constant, the performance of a non-adaptive filter would be degraded. If the change in variance was great enough, an adaptive filter would be required.

A good method of estimating a time-varying mean function could be applied in numerous areas, such as any sort of traffic or flow control problem, perhaps in quality control of large-batch or flow manufacturing processes, and any application where it is desirable to detect a change in the

process. The sensitivity of the filter is directly adjustable by the modeller through the noise covariance matrices Q and R.

A particularly useful application would be to the estimation of the rate parameter of a non-time-homogeneous Poisson process. If this can be done accurately, the process can be transformed to a stationary one [ref.6], which greatly expands the number of analytic tools that can be used.

The Poisson process is a counting process in continuous time, and to attempt to filter a string of interarrival time data would violate the sampling theorem. The times of arrival are the measurement times, and they are most certainly not made at equally spaced intervals. Instead, the filter may be designed to sample a counting process. At discrete intervals the filter would count the number of arrivals since some arbitrary time origin. If the process were to continue for a long time, the time origin might occasionally have to be reset to prevent computer overflow. It is easy to see how this sampling process could be implemented even if the input data were actually arrival instants in continuous time. The sampling interval should be small enough that there is low probability that more than one arrival would occur during a given measurement interval. Since the number of arrivals is monotone non-decreasing in time, a velocity or trend model would be appropriate. The input data would consist of integers. The state estimates would not. The non-integer estimate of number of arrivals up to the current time would not be useful to us. However, the

second element of the state vector, the velocity or trend, would in fact be the filtered arrival rate estimate. Since the process is noisy and non-Gaussian, a very low steady-state gain is appropriate.

The time-varying Poisson process cannot have constant variance, since the mean and the variance are equal. A low arrival rate implies high variance in the Poisson process, which is equivalent to high measurement noise, which requires low gain. A constant-gain filter would therefore be relatively more sensitive at low arrival rates than at high arrival rates. An adaptive filter could be easily designed to use the inverse of the rate estimate as the measurement noise variance estimate. Stability might require that the adjustment of the measurement noise variance be itself a filtering process, in which the incoming variance estimate is regarded as data.

C. MULTIPLE REGRESSION

If the regression constants are assumed known (or computed by other means) the design of an appropriate filter is quite straightforward, and quality of estimation is related directly to the quality of the model. Note that the velocity filter is simply the regression of velocity on position, where the slope parameter is known to be one. If the regression coefficients are assumed to vary in time, the problem becomes non-linear and is quite complex. Because of the immense applicability of this model, additional

developmental work is indeed a fertile field for future research.

D. SOME DESIGN CONSIDERATIONS

In applications where the noise covariance matrices R and Q, and the system dynamic model (state transition matrix Φ) are known or easily estimated, design is straightforward and has been successfully accomplished while remaining in the time domain. However, in applications where sweeping assumptions are required, a frequency-domain analysis could be very helpful. Some guidelines are as follows:

1. Spectral Analysis of the Data

A spectral analysis of sample data will show what the frequency response of the filter should be. The Fast Fourier Transform (FFT) program available in most computer libraries is generally easy to use. However, the FFT programs generally require an exact power of 2 for the number of data points. Hamming [ref.7] points out some pitfalls. Since stationarity is assumed, the data should be considered as a rotating cylinder, and if the starting and ending values are not similar, a discontinuity will exist in the spectrum. The data can be tapered and padded with zeros, but exactly the best method to accomplish this is unknown. Several methods might be tried.

The main virtue of the FFT is its speed. It works well on a long run of data. If the number of data points is small

(around a hundred) it might be effective to find (or write) a less efficient, conventional discrete Fourier transform program, which would not require padded, truncated, or tapered data if the starting and ending values are similar. If the FFT program used does not require an exact power of two for the number of input data points, it would be well to find out why not. The program may be doing the padding and tapering itself, and the analyst should be curious as to how. The analyst should remember that the spectrum is computed from the data, and it is therefore an estimate. If the run of data is short, there will be considerable variance in the estimate.

2. Frequency Analysis of Proposed Models

The analyst may test the effect of assumptions made in designing the filter by simply obtaining an impulse response of the filter and running it through an FFT. Truncation and tapering is no problem, because the impulse response will approach zero with time. The proper impulse function is simply a 1 followed by 2^n-1 zeros for a filter, or a 1 in the middle of 2^n-1 zeros for a smoother. If the output of the FFT consists of real and imaginary components, it will be necessary to compute amplitude and phase.

3. Adjusting the Model

If the model dynamics seem adequate but the bandwidth is wrong, the analyst should by now have some insight into what adjustments to make to the noise covariance matrices to

try to improve things. In a model of any complexity at all, there are numerous possible combinations. However, even some improvement over the initial assumptions will be beneficial. We are not looking for theoretical elegance, we are looking for performance.

Perhaps the model dynamics obviously call for a trend filter or even a change-of-trend (acceleration) filter, but the data is quite noisy. Consideration should be given to lowering the order of the filter. A very low-gain velocity filter will not follow changes in trend well. A higher-gain scalar filter may do so more effectively, although it will lag a steady trend. There are many tradeoffs, and we cannot achieve perfection.

4. Testing the Model

The model should be tested on real or simulated data. From here on, the modelling process is the standard cyclical one, going back to earlier steps as necessary until satisfactory performance is achieved.

APPENDIX A. DERIVATIONS

1. SCALAR KALMAN FILTER

a. Recursive Formula for Kalman Gain

The covariance extrapolation equation

$$P(t) = \Phi \Sigma(t-1) \Phi^T + Q$$

reduces in the scalar case to

$$P(t) = \Sigma(t-1) + Q$$

Since

$$K(t) = \Sigma(t) H^T R^{-1}$$

we may write, for the scalar case,

$$\Sigma(t-1) = K(t-1) R$$

Similarly, since

$$K(t) = PH^T [HPH^T + R]^{-1}$$

by reducing to the scalar case and substituting, we may write

$$K(t) = \frac{K(t-1) R + Q}{K(t-1) R + Q + R} = \frac{K(t-1) + Q/R}{K(t-1) + Q/R + 1}$$

b. Steady-State Kalman Gain

Rearranging the recursive gain equation and letting

$$K(t) = K(t-1) = K$$

we see that

$$K^2 + (Q/R)K - Q/R = 0$$

By the quadratic formula,

$$K = \frac{-Q}{2R} \pm \sqrt{\frac{Q^2}{4R^2} + \frac{Q}{R}}$$

We are oblised to take the larger root, since the smaller

root would force the gain to be negative. We also observe the inverse relationship

$$(K + Q/2R)^2 = Q^2/4R^2 + Q/R$$

$$K^2 + (Q/R)K + Q^2/4R^2 = Q^2/4R^2 + Q/R$$

$$K^2 = (Q/R)(1-K)$$

$$Q/R = K^2/(1-K)$$

c. Transient Kalman Gain

Recall that the Kalman filter requires a prior state estimate X(0) and a prior estimate of covariance P(0). This requirement can be avoided by using K(1) = 1, which allows the initial state estimate to be equal to the first measurement. Recall that

$$K(t) = \Sigma(t)\, H^T R^{-1}$$

Since K(1) = 1, then $\Sigma(1) = R$

d. Amplitude and Phase of Frequency Response

The frequency response is

$$H(v) = a \sum_{t=0}^{\infty} [b \exp(-iv)]^t$$

since

$$|b \exp(-iv)| < 1$$

then

$$H(v) = a \,/\, [1-b \exp(-iv)]$$

The amplitude squared is

$$A^2 = H(v)\, H(-v) = a^2 / [1-b \exp(-iv)]\, [1-b \exp(iv)]$$

By Euler's relation

$$A^2 = a^2 / [1 + b^2 - 2b \cos v]$$

which may be written

$$A^2 = a^2 / [(1-b)^2 + 2b(1-\cos v)]$$

recalling that

$$Q/R = K^2/(1-K) = a^2/b$$

we may write the amplitude as

$$A = \sqrt{\frac{Q/R}{Q/R + 2(1-\cos v)}}$$

The phase angle is

$$\theta(v) = \arctan [Im(v) / Re(v)]$$

where Im(v) and Re(v) are the imaginary and real parts of H(v), which may be written

$$H(v) = \frac{a \quad [1-b \exp(iv)]}{[1-b \exp(-iv)][1-b \exp(iv)]}$$

$$H(v) = \frac{a\ (1+b \cos v - ib \sin v)}{1 + b^2 - 2b \cos v}$$

which allows us to write

$$\theta(v) = \arctan [(-b \sin v) / (1-b \cos v)]$$

The angle for maximum phase shift occurs when

$$\frac{d\theta(v)}{dv} = \frac{b^2 - b \cos v}{1 + b^2 - 2b \cos v} = 0$$

so that the maximum phase shift $\theta(v)max$ occurs when

$$v = \arccos b$$

and has a value of

$$\theta(v)max = \arctan \frac{-b \sin(\arccos b)}{1-b \cos(\arccos b)}$$

$$= \arctan (-b / \sqrt{1-b^2})$$

2. IMPULSE RESPONSE FUNCTIONS

The impulse-response of the scalar Kalman filter is

$$g(t) = ab^t \qquad t = 0,1,2,\ldots$$

a. Impulse-response of Double Filter

$$g(t) = ab \circledast ab$$

$$g(t) = \sum_{k=-\infty}^{\infty} ab^k ab^{t-k} = a^2 \sum_{k=0}^{t} b^k b^{t-k}$$

$$g(t) = (t+1)\, a^2 b^t$$

b. Impulse-response of Scalar Kalman Smoother

$$g(t) = ab^t \circledast ab^{-t}$$

$$g(t) = \sum_{k=-\infty}^{\infty} a\, b^k ab^{t+k} = a^2 \sum_{k=0}^{\infty} b^{t+2k} = a^2 b^t \sum_{k=0}^{\infty} b^{2k}$$

$$g(t) = a^2 b^t / (1-b^2) = a\, b^t/(1+b)$$

BIBLIOGRAPHY

1. Kalman, R.E., "A New Approach to Linear Filtering and Prediction Problems", Journal of Basic Engineering(ASME), Vol 82D, pp 35-45, March, 1960.

2. Kalman, R.E. and Bucy, R., "New Results in Linear Filtering and Prediction", Journal of Basic Engineering(ASME), Vol 83D, pp 95-108, 1961.

3. Kailath, T., "A View of Three Decades of Linear Filtering Theory", IEEE Transactions on Information Theory, Vol. IT-20, pp 145-181, March, 1974.

4. Gelb, A., and others, Applied Optimal Estimation, The M.I.T. Press, 1974.

5. Naval Surface Weapons Center report NWSC/DL TR-3445, Development of an Adaptive Kalman Target Tracking Filter and Predictor for Fire Control Applications, by B.L. Clark, March, 1977.

6. Larson, H.J. and Shubert, B.O., Probabilistic Models in Engineering Sciences, Vol. 2, John Wiley and Sons, 1979.

7. Hamming, R.W., Digital Filters, Prentice-Hall, 1977.

8. Bloomfield, P., Fourier Analysis of Time Series: an Introduction, John Wiley and Sons, 1968.

9. Brillinger, D.R., Time Series; Data Analysis and Theory, Holt, Rinehart, and Winston, 1975.

10. Brown, R.G., Smoothing, Forecasting, and Prediction of Discrete Time Series, Arthur D. Little, 1963.

11. Box, G.E.P. and Jenkins, G.M., Time Series Analysis, Forecasting and Control, Holden-Day, 1976.

12. Reddick, H.W. and Miller, F.H., Advanced Mathematics for Engineers, 3rd ed., John Wiley and Sons, 1955.

13. U.S. Naval Postgraduate School Technical Report/Research Paper no. 72, Some Remarks on Exponential Smoothing, by P.W. Zehna, December, 1966.

14. Bessler, S.A. and Zehna, P.W., "An Application of Servomechanisms to Inventory", Naval Research Logistics Quarterly, Vol. 15, pp 157-168, June, 1968.

15. Harris, F.J., "On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform", Proceedings of The IEEE, Vol. 66, pp 51-83, January, 1978.

16. Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, 1977.

17. McWhorter, A., "Time Series Forecasting Using the Kalman Filter: An Empirical Study", ASA Proceedings of the Business and Economic Statistics Section, pp 436-441, 1975.

18. Downing, D.J., Pike, D.H. and Morrison, G.W., "Application of the Kalman Filter to Inventory Control", Technometrics, Vol. 22, pp 17-22, February, 1980.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Library, Code 0142 Naval Postgraduate School Monterey, California 93940	2
2.	Department Chairman, Code 55 Department of Operations Research Naval Postgraduate School Monterey, California 93940	1
3.	Professor R.W. Hamming, Code 52Hg Department of Computer Science Naval Postgraduate School Monterey, California 93940	1
4.	Professor D.P. Gaver, Code 55Gv Department of Operations Research Naval Postgraduate School Monterey, California 93940	2
5.	LCDR C.F. Taylor, Code 55Ta Department of Operations Research Naval Postgraduate School Monterey, California 93940	1
6.	LTCOL W.J. Costello, USMC 120 Hesketh Street Chevy Chase, Maryland 20015	1
7.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2